CHAPTER 2
LITERATURE REVIEW

2.1 CHAPTER INTRODUCTION

An overview of this thesis was given in the previous chapter. Chapter 2 covers the literature survey to provide the background knowledge for the research conducted. Figure 2.1 illustrates the flow of Chapter 2. This chapter starts with an introduction to digital image processing (Section 2.2). In this section, three topics are focused:- image processing (Section 2.2.1), computer vision (Section 2.2.2) and computer graphics (Section 2.2.3). Section 2.3 describes the technology of magnetic resonance imaging (MRI) and other issues related to MR images. The next section, Section 2.4, gives a comprehensive look on 3D imaging pipeline by Vannier *et. al.* in [2]. The steps in the pipeline are explored to show the relevance of this research to the bio-medical field. The different biomedical image segmentation techniques are presented in detail in Section 2.5. Other segmentation issues such as segmentation validation and applications are also discussed in this section. The next two sections talk about the techniques proposed to segment the MR images using artificial neural networks (Section 2.6) and fuzzy logic (Section 2.7). Section 2.8 summarizes this chapter.

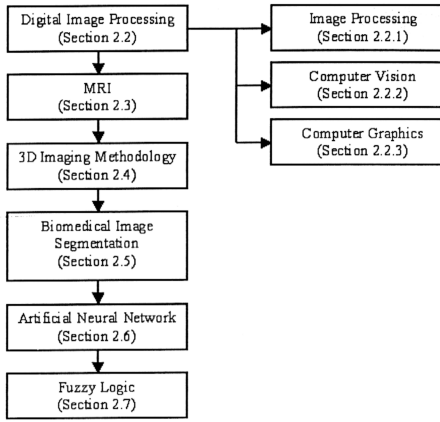


Figure 2.1 – Overview of Topic Discussed in Chapter 2

2.2 DIGITAL IMAGE PROCESSING

Computer imaging can be defined as the acquisition and processing of visual information by computers [1]. Computer imaging has roots in three additional field of computer science, which are image processing, computer vision and computer graphics [2]. Image processing entails with any image-to-image transformation [2]. Computer vision or image understanding is the construction of symbolic description from input images [1,2]. In computer vision applications the output images are for use of a computer, whereas the output of image processing applications are for human consumption. Computer graphics refer to the reproduction of visual data

through the use of computer [1,2]. These three categories are not entirely separate or distinct because the boundaries that separate them are nebulous.

2.2.1 IMAGE PROCESSING

Digital image processing encompasses of a broad range of hardware, software and theoretical knowledge. The fundamental steps in image processing are shown in Figure 2.2.

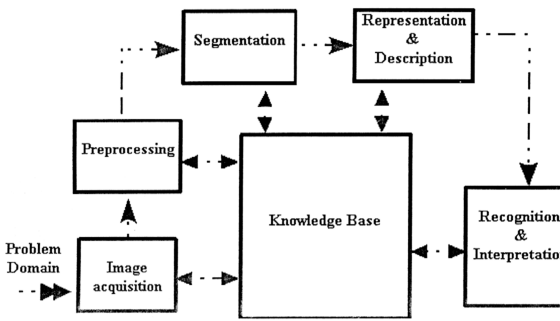


Figure 2.2 – Fundamental Steps in Image Processing [3]

The initial step in image processing is to acquire the digital image. In the domain of medical imaging, the subject is scanned using the medical modality available. Then the image is preprocessed to improve the image in ways to increase the chances of other processes. Noise removal and region isolation are some of the preprocessing techniques used. The next step is to segment an input image into constituent parts or

objects. Autonomous segmentation is the most difficult tasks in digital image processing [3]. The result of segmentation is usually raw pixel data, constituting either the boundary of a region or all the point in the region itself. The representation of segmented data is an important part for transforming raw data into a suitable form for further processing. Description or feature selection deals with the extraction of features that are basic for differentiating one class of object from another. The final stage is the recognition and interpretation of the image. Recognition is the process that assigns a label to an object based on prior knowledge. Interpretation involves assigning meaning to an ensemble of recognized objects.

Knowledge database contain information or data about a domain that is coded into the image processing system. The complexity of the knowledge base depends on the application of the image processing system. It guides the operations of each module and controls the interaction between modules. The results of image processing can be viewed in any stage of processing shown in Figure 2.2. Some system may not require all the steps shown in Figure 2.2. For example, image enhancement for human visual interpretation stops at the preprocessing step.

Major focuses in image processing are image restoration, image enhancement, and image compression. Image restoration is the process of taking an image with some degradation and restoring it to its original appearance [1]. Image enhancement involves taking an image and improving it visually by taking advantage of human visual system's responds [1]. One of the simplest enhancement techniques is to stretch the contrast of an image. Enhancement methods are problem specific. For

example, the enhancement method used to enhance medical images may not be suitable for enhancing satellite images. Image compression involves reducing massive amount of data needed to represent an image [1].

Image processing applications are used widely in the medical field. Diagnostic imaging modalities allow the medical professional to see the inside of a human body without having to cut it open. Image processing is used in different types of biological research, such as to enhance microscopic images to bring out features that are undiscernible.

2.2.2 COMPUTER VISION

Computer vision is computer imaging where the applications do not involve a human being. The images are examined by and acted upon by a computer. The final results of computer vision systems require a computer to use the visual information directly. Image analysis is a major area in computer vision.

Image analysis involves the examination of the image data to facilitate solving a vision problem [1]. Image analysis has two major focuses: - feature extraction and pattern classification. Feature extraction is the process of acquiring higher-level image information [1]. Pattern classification is the act of taking the high-level information and identifying objects within the image [1].

Computer vision systems are used in various types of environments. It is widely used within the medical environment. Current examples of medical systems are

- Automatic skin tumor diagnose
- Neurosurgery aid system during brain surgery
- Automatic clinical testing systems.

Automatic systems for diagnostic process are developed for the usage of medical professionals when specialists are unavailable and also as training tools for medical professionals. Computer vision systems used during surgery have improved the surgeon's ability to "see" and improve the quality of medical care. Computer systems are also used for tissue and cell analysis such as automatically identifying and counting certain type of cells [1].

2.2.3 COMPUTER GRAPHICS

Computer graphics is used to synthesize images from numerical descriptions [2]. Techniques were developed for realistic display of human defined object such as computer aided design models. Models are usually represented by infinite thin patches (triangles) or higher order curves (splines) [2]. Computer graphics contributed data structures (Section 2.4.1), projections techniques (Section 2.4.4) and shadowing techniques (Section 2.4.3) for use in 3D imaging [2,4]. Typically computer graphics techniques are used with image processing techniques [4]. They are used to model and study physical functions, design artificial limbs and, plan and practice surgery [4].

2.3 MAGNETIC RESONANCE IMAGING (MRI)

Figure 2.3 illustrates the flow of Section 2.3. A general idea about different medical imaging techniques is given in Section 2.3.1. Then issues pertaining MRI is discussed in detail as illustrated in Figure 2.3. Finally in Section 2.3.11, multi modality registration is discussed in short.

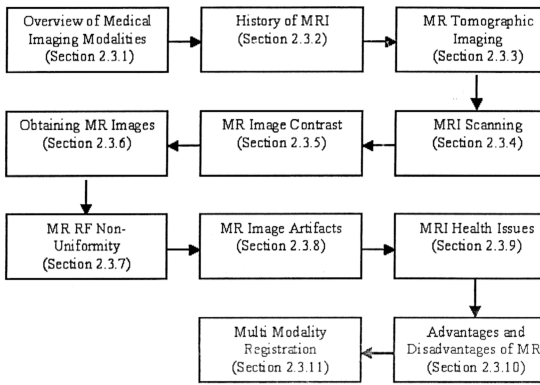


Figure 2.3 – Overview of Section 2.3

2.3.1 OVERVIEW OF MEDICAL IMAGING MODALITIES

Several kinds of non-invasive digital medical imaging techniques are available for visualization of internal organs. Among the popular imaging methods used in biomedical application are :-

- ***X-ray Projection Imaging***

  X-ray is a form of radiant energy used in the diagnosis and treatment of illness. X-ray gives excellent bone rendition despite the fact that it does not provided in-depth information of an object [2]. X-rays are deflected and absorbed to different degrees by various anatomies in the body. The absorption amount depends on the tissue composition. Dense materials absorb more x-ray than soft tissues [7]. The first application of X-ray was to visualize bone fracture.

- ***Computed Tomography (CT) (1972)[8]***

  CT became available in the mid 1970's [7]. CT images are produced by measuring radiation along a circular span path [9,2]. CT scans are suitable to image high density objects such as bone [2]. 3D imaging is achieved by rotating the x-ray emitter around the patient and measuring the ray intensity from different angles [7].

- ***Ultrasonography (US) (1978)[8]***

  US uses high frequency sound waves to produce real time images of body organs [9] and does not involve ionizing radiation [10]. US produces a rather low image quality, the depicted structures are mainly borders between different organs [2]. It is a real-time tomographic imaging modality [10,7]. Furthermore, it also can be used to produce real-time images of tissues and blood flow velocity [8,7]. US is portable (US examination can be performed at bedside) and is relatively low cost [10].

- ***Positron Emission Tomograpy (PET)***

  PET uses positron decay patterns to study metabolic reaction of the human body systems [9,2]. PET's purpose is to study regional metabolism and neuroreceptor activity in the brain and other tissues [7]. Its clinical usage is in cancer detection of the brain, breast, heart, lung and colorectal tumor [7]. Research conducted using PET includes the study of epilepsy, brain tumor, stroke and Alzheimer's disease.

- ***Single Photon Emission Computed Tomography (SPECT)***

  SPECT is used to detect metabolic activities of the real world object [9,2,7]. SPECT is inferior to PET due to the attainable resolutions and sensitivity [7]. Although SPECT is not as good as PET, the availability of SPECT, the conveniences and economical aspects of SPECT makes is attractive for clinical studies of the brain.

- ***Electrical Source Imaging (ESI)***

  ESI is a technique used for reconstructing electrical activity in the brain or heart [7]. ESI records abnormal electrical activity of the heart which allows doctors to localize the abnormal electrical pathways and treat it [7].

- ***Electrical Impedance Tomography (EIT)***

  EIT uses low frequency electrical current to construct a map of the conductivity or resistively of the region of the body [7].

- ***Magnetic Source Imaging (MSI)***

  MSI is the general term used for the reconstruction of current sources in the heart or brain from the measurements of external magnetic fields [7].

- ***Laser Optical Tomography (LOT)***

  LOT reconstructs the amount of transmitted laser light through the body along multiple paths [7]. It is used to measure tissue oxygenation for muscular dystrophy, tissue perfusion in diabetic disease, the detection of brain hemorrhaging, monitoring stroke patients, and the study of brain activity during specific tasks [7].

- ***Magnetic Resonance Imaging (MRI) (1978)[8]***

  MRI is an imaging technique used in the medical field to produce images of the inside of the human body. It is based on the nuclear magnetic resonance (NMR), a technique used by scientist to obtain microscopic chemical and physical information about molecules. MRI scans are very sensitive to variations of soft tissues [2]. MRI is a tomographic imaging technique, since it produces images of NMR signals in slices. Today, MRI has advanced beyond tomographic imaging to volume imaging technique [11].

The utilization of diagnostic imaging technology has increased knowledge of normal tissue and diseased anatomy in medical research, and is a critical component in diagnostic and treatment planning.

Table 2.1 – MRI Timeline [11]

Magnetic Resonance Imaging (MRI) Timeline	
1946	MR phenomenon by Bloch & Percell
1950	NMR developed as analytical tool
1960	
1970	
1972	Computerized Tomography
1973	Backprojection MRI - Lauterbur
1975	Fourier Imaging - Ernst
1980	MRI demonstrated - Edelstein
1986	Gradient Echo Imaging NMR Microscope
1989	Echo-Planar Imaging
1991	Nobelr Prize - Ernst
1994	Hyperpolarized ^{129}Xe Imaging

2.3.3 MAGNETIC RESONANCE (MR) TOMOGRAPHIC IMAGING

MRI started out as tomographic imaging technique for producing NMR images of a slice of the human body, with a thickness. MR images are composed of picture elements or pixels. The intensity of a pixel is proportional to the NMR signal intensity of the contents of the corresponding voxel of the object being imaged [11]. MRI is based on the absorption and emission of energy in the radio-frequency range of the electromagnetic spectrum.

MRI takes advantage of the primary content of human body which is fat and water. Fat and water molecules contain many hydrogen atoms, which makes the human body about 63% hydrogen atoms [12]. This allows the MRI to image the NMR

signals from the hydrogen nuclei. The contrasts in MR images rely on the amount of hydrogen in the anatomy scanned. Fat and water produce bright pixels in the MR image while in contrast the bone produces dark pixels.

2.3.4 MAGNETIC RESONANCE IMAGING (MRI) SCANNING

When a patient undergoes MRI procedure (Figure 2.4), three major events transpire :-

1. First the patient is placed in a magnetic field
2. Then a radio-frequency pulse is applied
3. Then the pulse is terminated to allow relaxation to occur [12]

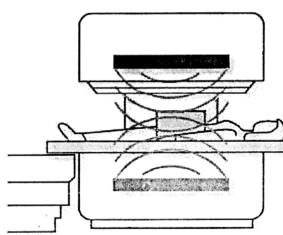


Figure 2.4 – MRI Scanner

Magnetic Field - A hydrogen atom contains a single proton that has a property called spin or angular momentum [13,12,11], which will cause the nucleus to produce an NMR signal. The moving charge also produces its own magnetic field (Figure 2.5(a)). So each hydrogen nucleus is like a tiny magnet bar. When placed in an

external magnetic field, the protons will precess at an axis parallel to the magnetic field as shown in Figure 2.5(b) [13,14]. The precession rate of a proton is directly proportional to the strength of the magnetic field [14]. The precession frequency is called "Larmaor Frequency". The external magnetic field used in MRI is between 0.5 to 1.5 Telsa [12].

Radio-frequency Pulse - A radio-frequency pulse is an electromagnetic wave that results from the brief application of an alternating current [12]. Radio frequency pulse is applied perpendicular to the magnetic field for a brief period of time and then shut off. The frequency of the radio-frequency pulse is tuned to the Larmoaur frequency. This concept is called "resonance". The proton will absorb the radio-frequency energy and alter its precession as in Figure 2.5(c) [14,13].

Relaxation - Relaxation is the process that occurs after the termination of the radio-frequency pulse, in which the physical charges that were caused by radio-frequency pulse return to the state of equilibrium. The system would release the energy in the form of radio frequency as shown in Figure 2.5(d). Different tissues have different rates of relaxation, thus different signal intensities are obtained, which produced tissue contrast. Relaxation of the protons to their equilibrium state has two nuclear magnetization components [12,14,13]:-

1. Component which is parallel to the external magnetic field is known as time T1 or longitudinal relaxation time
2. Component which is perpendicular to the external magnetic field is known as time T2 or transverse relaxation time.

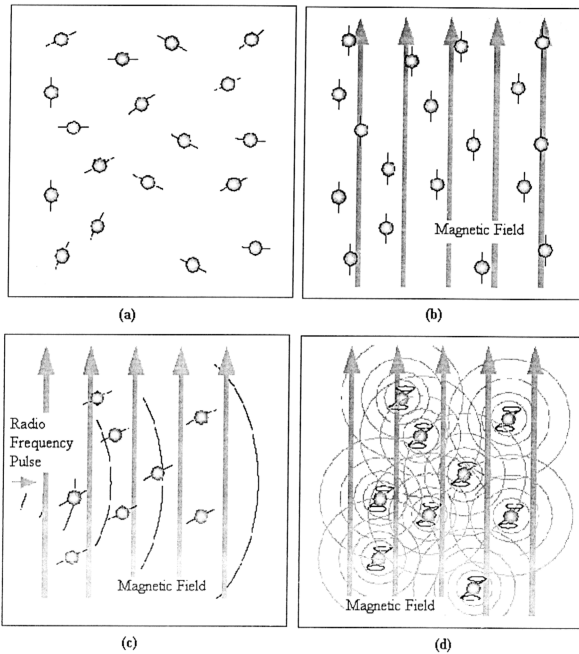


Figure 2.5 - Hydrogen Atoms During MRI scanning [15].

(a) Hydrogen atoms in their natural state before the external magnetic field is applied

(b) Precession of the hydrogen atoms after the introduction of the magnetic field

(c) Hydrogen atom when the radio-frequency pulse is applied

(d) Hydrogen atoms after the radio-frequency pulse is terminated

Repetition time, TR, is the period when the radio-frequency pulse sequence is repeated at a pre-determined rate (Figure 2.6) [12,11]. The energy released or free induction decay is measured at various times within the TR interval. The echo decay time, TE is the time between which the radio-frequency pulse is applied and the release signal is measured (Figure 2.6) [12,11]. T1 and T2 refer to tissue properties while TR and TE refer to equipment parameter [12]. By manipulating TR and TE, different signal strength is obtained to produce tissue contrast [12,11].

2.3.5 MAGNETIC RESONANCE (MR) IMAGE CONTRAST

The MRI signal strength depends on these parameters [16,14,15] :

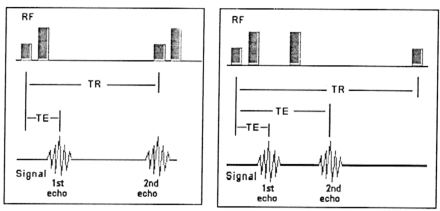
1. Proton density – concentration of protons in the form of water or fat in tissue.
2. T1 relaxation time
3. T2 relaxation time

Contrast on MR image is manipulated by changing the pulse sequence parameters which are [16,12]:-

1. The specific number, strength, and timing of the radio-frequency pulses.
2. The specific number, strength, and timing of the gradient pulse.
3. TE (Echo Decay Time)
4. TR (Repetition Time)

Common pulse sequences are [16,12]:-

1. T1 weighted spin-echo sequence – short TR (<1000 ms) and short TE (<30ms) (Figure 2.6a).
2. T2 weighted spin-echo sequence – long TR (>1500 ms) and long TE (>60ms). T2 weighted sequence is employed as dual echo sequence. A first shorter echo (TE<30ms) is weighted (Figure 2.6b).



(a) T1 weighted **(b)Dual Echo T2 weighted**

Figure 2.6 – Spin Echo Pulse Sequence

Tabulation of tissue image characteristics is shown in Table 2.2 [14]:

Table 2.2 - Tissue of Image Characteristics

	T1 weighted	T2 weighted
Bone	dark	dark
Air	dark	dark
Fat	bright	bright
Water	dark	bright

2.3.6 OBTAINING MAGNETIC RESONANCE (MR) IMAGES

A magnetic resonance image acquisition consists of a set of planar images through some volume of tissue [12]. Each plane (slice) contains a two dimensional image which is composed of intensities located on this plane with the use of two coordinates. Portrayal of the image in each plane requires two coordinates axes. A third axis is necessary to locate the position of each plane where there are multiple planes situates at different levels [12]. Magnetic resonance image acquisition requires three coordinates axes due to its three-dimensional phenomena [12].

Since there are three coordinates to localize a point in space, three magnetic field gradients oriented along the three major axes are required. One gradient is used to locate the level of each place is known as the "slice select" gradient. Two other gradient are used to locate points of intensity within each plane, are called "frequency" and "phase encoding" gradient [12]. The slice select gradient will be the z-axis, and the x and y axes will represent the frequency and phase encoding gradient receptively. 2-D Foureir Transform is used to transform the encoded image to spatial domain.

MRI scan is performed on biological objects and the images produced use the anatomical coordinate system, where the axes are referenced to the body. The z-axis runs along longitudinally parallel in the head-foot or superior-inferior (S/I) direction, the x-axis runs parallel in the left-right (L/R) direction and the y-axis runs parallel in the front-back or anterior-posterior (A/P) direction [11]. The three axes are illustrated

in Figure 2.7.

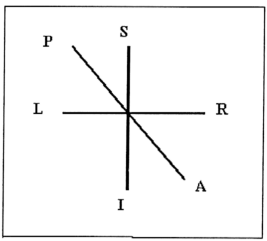


Figure 2.7 - Axes Referenced to the Body

Three planes are formed by anatomical coordinates system :-

1. ***Transverse*** – The L/R and A/P axes form the transverse plane, which is imaged perpendicular to the long axis of the body.
2. ***Coronal*** – The L/R and S/I axes form the coronal plane, which bisects the body from the back to the front.
3. ***Sagittal*** - The S/I and A/P axes form the sagittal plane, which bisects the body from the left to the right [11] .

Figure 2.8 illustrates the three planes.

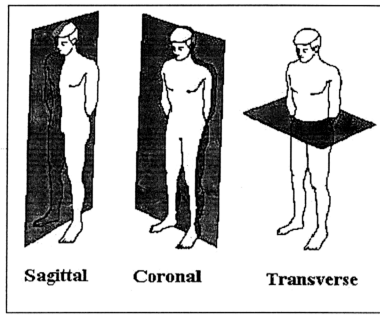


Figure 2.8 - Sagittal, Coronal and Transverse Planes

Images obtained from the MRI scanner are a set of images in the direction of any given plane. For example, when the leg of a patient is scanned in the direction of the transverse place, a set of transverse images of the MRI scan would be obtained. Usually scans are not performed in the direction of all three planes due to cost and time constrains. Since traditional use of MR images were used by medical professional such as doctors, they are trained to diagnose from the given set of images.

Transverse, coronal and sagittal slices are usually obtained for computer graphic manipulation and modeling of volumetric data. All three sets of images are required

to capture more information and measurements that would be an aid in the computer field. The three sets of images are aligned to obtain volumetric or 3D images. The volumetric or 3D images are represented by voxels. Voxels or volume elements are referenced by x, y, and z coordinates [11].

The depth of a voxel is set by the thickness of the slice of an anatomical structure that is being modeled. The width and height are determined by the :

1. area that within the scope or boundaries of the MRI scan.
2. overall size of the image
3. number of the columns or rows of voxels that are used to generate the images.

Each voxel is represented by black, white or shades of gray. The shade, or intensity, of the voxel is determined by the duration of the radio-frequency response that is emitted from the corresponding area of the patient's anatomy.

2.3.7 MAGNETIC RESONANCE (MR) RADIO FREQUENCY (RF) NON-UNIFORMITY

MR image non-uniformity may be caused by a number of factor [17]:-

- RF transmitter and receiver inhomogeneities
- RF penetration
- Static field inhomogeneity

Segmentation accuracy depends on the quality of the data and the MRI data limitation is the image non-uniformity due to RF field inhomogeneity. The correction methods proposed fall into general frame work of [17]:-

1. ***Correction Matrix*** – In this method, the coil sensitivity profile is found using a phantom containing a uniform solution of MR contrast material, which is then divided into the original material. However, this method adds spatial dependence to the noise. This method also assumes that RF penetration through the phantom and the same subject is the same, which is not the case [18].

2. ***Digital filtering*** - A digital filter is derived on the assumption that the true image is comprised of smooth background with high frequency details superimposed. This method may not be used as universal correction as the construction premise is not totally right.

3. ***Surface fitting*** - Curve fitting routines have been used to approximate phantoms image pixels values as a function of their position relative to some reference value. This is then used to construct a correction matrix. Results indicate that this is an effective method.

4. ***Statistical method*** - This approach uses knowledge of the tissue properties and intensity inhomogeneities to correct the RF non-uniformity.

5. ***Maximum entropy method (MEM)*** – Contradicting finding exist about the use of MEM for noise removal. Constable *et la.*[17] reports that MEM did not produce satisfactory results. However, Moran *et la.* [17] indicates that MEM has strong potential for further application in MR.

RF uniformity correction depends upon the patient, slice orientation, RF coil design, pulse sequence and the method of choice applied to the subject image [17]. Most of the methods were designed for the use with MRI brain scans, where tissue contrast is obvious. These methods will not work for femur scans because of intensity difference is very small between femur, cartilage and soft tissues [19]. When the methods are applied, information may accidentally be removed [20]. Studies on intensity correction are done to assist the worst cases, so when applied to less affect data, it may induce more inhomogeneities [21].

2.3.8 MAGNETIC RESONANCE (MR) IMAGE ARTIFACTS

Image artifact is any feature that is not present in the original object, but appears in the scanned image. It is sometimes the result of improper operation of the scanner, and sometimes due to natural process or properties of the human anatomy.

Table 2.3 gives a summary of a few MRI artifacts [11].

Table 2.3 - Summary of MR Image Artifacts

Artifact	*Cause*
RF Quadrature	Failure of the RF detection circuitry
Magnetic Field Inhomogeneity	Metal object distorting the magnetic field
Gradient	Failure in a magnetic field gradient
RF Inhomogeneity	Failure of RF coil
Motion	Movement of the imaged object during the sequence
Flow	Movement of body fluids during the sequence
Chemical Shift	Large magnetic field and chemical shift difference between tissues
Partial Volume	Large voxel size
Wrap Around	Improperly chosen field of view
Gibbs Ringing	Incomplete digitization of the echo

1. ***RF Quadrature*** - RF quadrate artifacts are caused by problems with the RF detection circuitry. This artifact is the result of a hardware failure and must be addressed by a service representative [11,22].

2. ***Magnetic Field Inhomogeneity*** – An inhomogeneous magnetic field will cause either spatial, intensity or spatial distorted images [11,22].

3. ***Gradient*** - An gradient which is not constant with respect to the gradient direction will distort an image. This is typically only possible if a gradient coil has been damaged [11].

4. ***RF Inhomogeneity*** - RF inhomogeneity is caused by either a nonuniform magnetic field or an nonuniform sensitivity in a receive only coil. The presence of this artifact represents the failure of an element in the RF coil or the presence of non-ferromagnetic material in the imaged object. Figure 2.9 shows a sagittal image of the head containing RF inhomogeneity artifact in the mouth region due to large amount of non-ferromagnetic dental work [11].

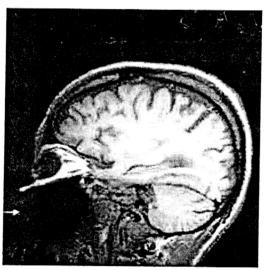


Figure 2.9 – RF Inhomogeneity Artifact (illustrated by the arrow)

5. ***Motion*** - Motion artifacts are caused by motion of the imaged object or a part of the imaged object during the imaging sequence. Figure 2.10 shows a motion artifact due to the movement of a blood vessel. This motion caused a ghosting across the image [11,12].

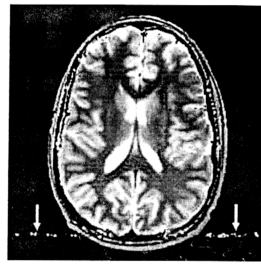


Figure 2.10 - Motion Artifact (illustrated by the arrows)

6. ***Flow*** - Flow artifacts are caused by flowing blood or fluids in the body. Figure 2.11(a) shows an example of flow artifact from an axial slice through the legs, where the blood vessels appear black even though they contain large amount of water [11].

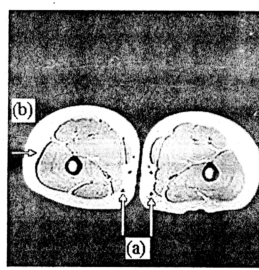


Figure 2.11 – MRI Artifacts (a) Flow Artifact (b)Chemical Shift Artifact

7. ***Chemical Shift*** - A chemical shift artifact is caused by the difference in chemical shift (Larmor frequency) of fat and water. The artifact manifests itself as a mis-registration between the fat and water pixels in an image. Figure 2.11(b) shows chemical shift artifact between the fat and muscle in the legs [11,22,12].

8. ***Partial Volume*** - A partial volume artifact is any artifact which is caused by the size of the image voxel. For example, a larger voxel might contain a combination fat and water so the signal intensity of that voxel equals to the weighted average of the quantity of water and fat present in the voxel. This artifacts will also manifestation a loss of resolution caused by multiple features present in the image voxel [11].

9. ***Wrap Around*** - A wrap around artifact is the occurrence of a part of the imaged anatomy, which is located outside of the field of view, is inside of the field of view. This artifact is caused by the selected field of view being smaller than the size of the imaged object [11,22].

10. ***Gibbs Ringing*** - Gibbs ringing is a series of lines parallel to a sharp intensity edge in an image which is caused by incomplete digitization of the echo [11,22].

2.3.9 MAGNETIC RESONANCE IMAGING (MRI) HEALTH ISSUES

MRI imaging is used widely in the medical field because there is no safety concerning X-irradiation [22,23]. There is a possibility that magnetic field could induce electric current that could cause ventricular fibrillation, cardiac arrest, inhibition of respiration or muscular respiration, though these symptoms were not observed in the England [23]. Studies in England only reported about 3% of patients who underwent MRI scanning have development a feeling of claustrophobia when positioned in the magnetic field [23]. The good record of MRI scanning is attributed to the caution exercised by radiographers. This is because patients with cardiac pacemakers, large prices of metal, metallic cranial aneurysm clip or metallic intrauterine devices, pregnant women and people with certain medical conditions such as epilepsy have been excluded from MRI scanning [23,22]. Most health hazard studies conducted use electromagnetic frequency ranging from 13MHz to 20,000 MHz while MRI magnetic field uses frequency ranging from 1MHz to 70MHz [24].

There is a danger posed from MRI scans, when there are ferromagnetic objects present. A ferromagnetic object becomes a projectile if it is in a magnetic field. Patients are checked for ornaments such as keys, pens, belts, or metal on clothing before foregoing MRI scans [22].

[24] has reports of MRI examinations. Cases include

- death which happened during a cerebral infarction examination.
- two cardiac arrests;
- pain in a patient with neuro-stimulator.
- eye or peri-orbital irritation due to ferromagnetic ingredients in eye make-up

2.3.10 ADVANTAGES AND DISADVANTAGES OF MAGNETIC RESONANCE IMAGING (MRI)

Advantages of MRI are :-

- It does not damage living tissue, so it is more acceptable for human experimentation [24,25]
- It can image objects that are transparent to light.
- It distinguishes soft and hard tissue well providing reasonably high resolution volumetric images.
- It is able to produce multi-planar images without moving the patient, thus reducing motion artifacts in slices produced [26,25].
- It has short examination time that reduces patient discomfort.
- It provides large anatomic regions without exposing patients to radiation compared to CT.
- MRI images are readily available for use.
- Its acquisition speed is faster than CT [8]
- It measures chemical properties uniformly.

MRI also has its down side:-

- Patients with metallic implant such as orthopedic implants and pacemaker cannot be sent for MRI scanning [25].
- Patients can be too big to be scanned
- Patients who are claustrophobic cannot be sent for MRI scanning
- The MRI scanner makes a lot noise due to the rising electrical current in the wires of the gradient magnets being opposed by main magnetic field [25]. Patients are usually given ear plug during scanning.
- Patients are required to hold very still for an extended period of time to reduce motion artifacts [25].
- MRI systems are very expensive, and therefore the MRI examination is also expensive [25].
- MRI scanner is not portable. Scanning can only be done in a special MRI room.

2.3.11 MULTI-MODALITY REGISTRATION

Multi modal imaging attempts to take advantage of the different kinds of anatomical provided by different imaging modalities. The multi modalities show different, complementary and/or partially overlapping aspects of the anatomy examined [27]. The use of multi modal depends on the availability of the images. However, even when the data is available, they may not be in the same alignment and require registration [28].

Registration techniques have developed for tomographic three and four dimensional

analyses of human perception and cognitive functions. Registration is the correlation of spatially related images [27,17]. It is the central process in the developing aspects of medical imaging [17]. Registration is a multistep procedure where the actual alignment of spatial data is just a component [27,17]. In some instances, segmentation is a prerequisite for registration, while in other instances, segmentation is performed on multi-spectral data that require registration [17].

Image registration combines the spatial information from modalities like MRI and CT, with functional data from SPECT or PET [8]. It can also take advantage of the bone imaging in CT and the soft tissue imaging capabilities in MRI. Registration combines images such that the strengths of the combined methods summate while the weaknesses canceled. Images registered can be [17] :-

- Intra-subject – from a single patient, either from multi-modality scans or from single modality scans
- Inter-subject – from several patients from a single modality

The image registration techniques are as follows :-

1. Physical head restraint [17]
2. External fiducial markers [27,17]
3. Surface matching [17]
4. Principle axes [17]
5. Internal anatomical land markers [17]
6. Mutual information [27]

The first two methods are prospective in nature, because it requires rigidly controlled setting when acquiring data [27,17]. Markers are especially made for good visibility and ease automatic detection in images [27]. This method is also patient unfriendly [27]. The next three methods are retrospective techniques, because of the ability to use previous acquired data and not require added control in the acquisition phase. They rely on being able to identify corresponding features on the images to be registered and calculating the transformation to match these features. Mutual information is best and most accurate multi-modality matching measure [27] between gray level values. It is a measure from the field of information theory introduced in the medical field in 1995.

2.4 3D IMAGING METHODOLOGY

Figure 2.12 presents an overview of 3D imaging methodology.

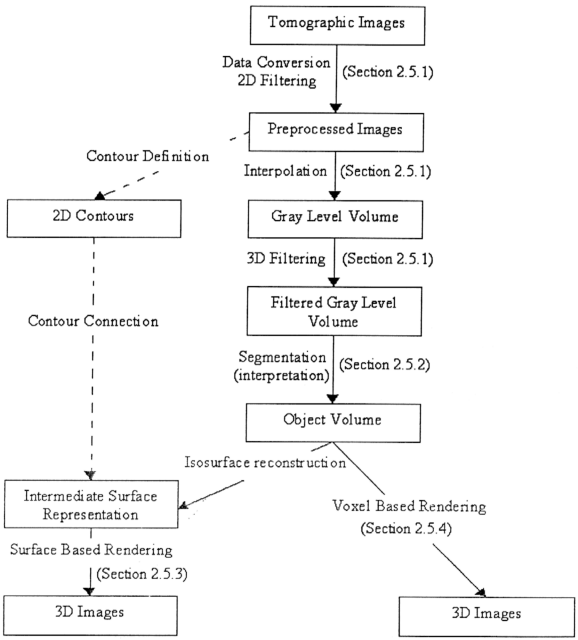


Figure 2.12 : General Sketch of Imaging Pipeline [2]

After a patient is scanned, a series of parallel cross sections are obtained. The data then usually undergoes some preprocessing for data conversion and image filtering. From here, the data processed can be either 2 dimensional (2D) contours (represented by dotted line) or volume data (represented by solid line).

The dotted line in Figure 2.12 represents an approach where an object is reconstructed from its contours on the cross-sectional images. The solid line represents the processing of contiguous data volume. Interpolation is used to achieve equal spacing in all three directions in the volume data. Filtering also can be performed to improve image quality and suppress unwanted information [99] (Section 2.4.1).

Identifying different objects represented in the data volume is done next. This step allows the object to be removed or selected for visualization. This is done through segmentation and interpretation (Section 2.4.2). Threshold is the simplest segmentation method used in CT data to distinguish bone from other data. However, MRI data require more sophisticated segmentation method. This is the field of research in this project.

The next step is rendering the segmented object. There are a few choices of which rendering technique use. Surface based rendering uses any standard computer graphics method (Section 2.4.3). It first creates an intermediate surface representation of the object shown. Voxel based methods were developed more recently (Section 2.4.4). It creates a 3 dimensional (3D) view directly from the

volume data. Volume visualization is a method of extracting meaningful information from volumetric data sets through the use of interactive graphics and imaging [29]. Volume rendering provided physicians with more data, giving them the confidence to operate and improve patient's quality of life [30]. It uses the full gray level information to render surfaces, cuts, or transparent and semi-transparent volume objects [31]. The decision of which rendering technique to use depends both on the available memory and computing power and the visualization goals. Volume graphics has advantages over surface graphics. It is viewpoint independent, insensitive to scene and object complexities, and suitable for representing sampled and simulated data sets [29].

2.4.1 PREPROCESSING

Data Conversion - During data conversion, changing data format may be required and reduction of data to save storage space and processing time. Common technique for data reduction are :-

1. Cutting : selecting region of interest
2. Reduced spatial resolution : matrix size is reduced. e.g. by averaging from 1024 pixels to 512 pixel
3. Reduced intensity resolution : e.g. from 32 bits to 16 bits.

Method 2 and 3 will generally cause loss of information.

Filtering - Image filtering is a general term for image processing routines that are used to smooth or enhance the image. An example is improving the signal-to-noise

ration in MRI images using image filters such as average, median, and Gaussian filters. But these filter also tend to smooth out small details in the images. Better results are achieved if 2D filters are used for images and 3D filters for volume. There are two types of filters :- suppressing filters and enhancing filters [99].

Interpolation - When 2D images are stacked on top of each other, a contiguous gray level volume is obtained. The resulting data structure is an orthogonal 3D array of voxels, called voxel model. Most algorithm for 3D imaging work on isotropic volumes where the sampling density is equal in all three dimensions. However, only a few data sets have this property. The missing information is reconstructed in the interpolation step [99]. Simple interpolation method is image replication. For better results, linear interpolation or splines are used to interpolate intensities between adjacent images. Interpolation is necessary when the objective is [99] :-

- To change the non-isotropic discretization of the input scene to isotopic discretization.
- To represent longitudinal scene acquisition in a registered common coordinate system
- To represent multimodality scene acquisition in a registered coordinate system.
- To re-section the given scene.

Data Structures - The most common data structures for volume data are :-

- Binary voxel model – a very simple model with voxel values are either 0 or 1.
- Gray level voxel model – in this model each voxel holds an intensity information
- Generalized voxel model – in this model, each voxel contains further information

such as the normal and gradient of the intensity, membership label or material percentages.

2.4.2 OBJECT DEFINITION

MRI images usually have large number of different of organs. Users decide which part of the data is needed and which is not. To display or manipulate the interested part, the computer has to know which voxel, for example is the bone and which is not. This information is also required for morphometric measurements such as distance, angles and volume. Object definition establishes relationship between voxels and anatomical terms. This task has two parts :- segmentation and interpretation. An object volume maybe organized using generalized voxel-model.

Segmentation - Segmentation is the process of partitioning gray level volume into different regions that are homogeneous with respect to some formal criteria and corresponding to real anatomical objects. Thresholding is the common segmentation technique used to select bone or soft tissue in CT images. Segmentation becomes more complicated if different organ with similar gray level characteristics are to be distinguished. Advanced segmentation methods researched for use in these cases are presented in Section 2.5.

Interpretation - An interpretation step is performed to identify various regions and labeled with meaningful terms such a bone or fat [28]. This step is skipped for simple applications. Labels are visually obvious to the users Computer automated

labeling is used when labels are not obvious and in automated processing.

2.4.3 SURFACE BASED RENDERING

Surface based rendering extracts an intermediate surface description of the relevant object from the volume data. This information is then used for rendering. This method reduces very high volume data to surface representation which reduces memory requirement and computer processing time. Nevertheless, surface reconstruction throws away an enormous amount of valuable information of the cross sectional images.

Surface Reconstruction from Contours - In this method, the object contour is defined in every tomographic image. Afterwards, the contours from adjacent cross-section are connected to form a 3D structure. In medical data, shapes are extremely complex and vary greatly from slice to slice, thus it is crucial to connect the different contours properly.

Surface Reconstruction from Volumes - This method creates isosurface representing all points at a certain intensity value found in the original gray level [31]. The surface can also be defined using object membership labels. Marching cubes algorithm is an algorithm developed recently which also utilizes gray level information.

Shading - Shading is the realistic display of an object based on position orientation, characteristics of its surface and the light sources illuminating it.

2.4.4 VOXEL BASED RENDERING

In voxel base rendering, no intermediate surface representation is required. Images are created directly from the volume data. Compared to surface based rendering, all original gray level information is maintained during rendering. Thus, this makes it an ideal technique for interactive data exploration. It also allows a combined display of different aspects such as opaque and semi-transparent surfaces, cuts, and maximum intensity projections. This technique processes large amount of data. Volume rendering also is effective for rendering real (measured) and numerical (calculated) data [30].

Projection Technique - In voxel-based rendering, there are two scanning strategies : image order (pixel by pixel) and volume order (voxel by voxel)

- **Image order** strategies scan the data volume on rays along the view direction. Ray casting is a common image order strategy used in voxel based rendering. Ray casting consists of tracing a lie perpendicular to the viewing plane into the scene domain [99]. Ray casting is a very flexible and intuitive scanning strategy that allow the integration of opaque, semi-transparent, and transparent display technique comparatively easy. But, ray casting algorithm is limited both by high memory and computing requirements.
- **Volume order** strategies scan the data volume along lines or columns of the 3D array. Volume data is traversed in back-to-front or front-to-back order.

Surfaces - After the projection methods are applied, thresholding or object membership label can be used to view the surface of an object. Then shading is applied for better visualization.

Cut Planes - Cutting is used to visualize interior structures after the surface view is available.

2.5 BIOMEDICAL IMAGE SEGMENTATION

Figure 2.13 illustrates the flow of Section 2.5. In this section numerous MR image segmentation methods are explored. Traditional methods such as thresholding (Section 2.5.1.1), edge-based approach (Section 2.5.1.2), region growing approach (Section 2.5.1.3), classifiers (Section 2.5.1.4) and clustering approach (Section 2.5.1.5) are discussed along with the advantages and disadvantages of the different methods are given. Researches on MR image segmentation using intelligent methods such as Markov random field model (Section 2.5.1.6), artificial neural networks (Section 2.5.1.7) and atlas guided approach (Section 2.5.1.8) are commended after that. Different segmentation validation methods are given in Section 2.5.2. Section 2.5.3 looks into the human femur, the anatomy of interest in this research. Previous segmentation of MR bone images are presented in Section 2.5.4. Finally, Section 2.5.5 explores the diverse application of medical image segmentation.

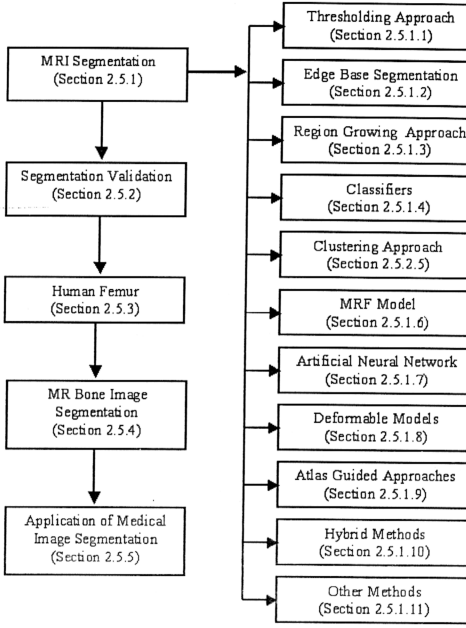


Figure 2.13 – Overview of Section 2.5

Segmentation is a wide spread technique used in image processing to reduce image information. Segmentation is the process of dividing or decomposing an image into mutual exclusive and meaningful regions using some algorithm so that each region has some properties that are uniform, homogenous, and differs from those of its

neighboring region [32,33,34,35,36,28,2]. Segmentation is usually a pre-processing step for other medical applications such as surgery planning, diagnostic, anatomical studies or biomodelling [37,8].

Anatomical structures are complex and non-symmetrical unlike man-made structures. These characteristics of anatomical structures make segmentation a challenging process [38]. There is also no standard model of a given anatomy due to the varying nature of anatomy from individual to individual. Variations are also due to heredity and deformations such as tumors, broken bones and deformations due to sickness. An example would the segmentation of brain tumors, where no two tumors are alike. It can be concluded that there are no one universal algorithm that best segment data for a specific application or an algorithm that solves all segmentation problem for any data [18,39,28]. There are some general methods that can be applied to a variety of data. Nevertheless, specialized methods for a particular application achieve better performance when aprior knowledge is taken into account.

The performance of segmentation method varies depending on the specific application, imaging modality and other factors. An example would be the segmentation of bone have different requirements from segmentation of brain tumor. Imaging artifacts such as noise, partial volume effect and motion also have consequences on the performance of segmentation algorithm.

Soft segmentation allows regions or classes to overlap [28]. It is important because of partial volume effect, where multiple tissues contribute to a single pixel or voxel

resulting in blurring of intensity across borders. Figure 2.14 shows how the sampling process can result in partial volume effect. In Figure 2.14(b), the acquired image, it is difficult to determine the boundaries of the two objects. Hard segmentation forces a decision if a pixel or voxel in inside or outside the object [28]. However, soft segmentation allows uncertainly in the location of the object boundaries. A characteristics function is an indicator function of whether a pixel is inside or outside of its corresponding set [28]. Characteristic functions can be generalized to membership functions which need not be binary valued. Soft segmentation based on membership functions can be easily converted to hard segmentation by assigning a pixel to its class with the highest membership value [28].

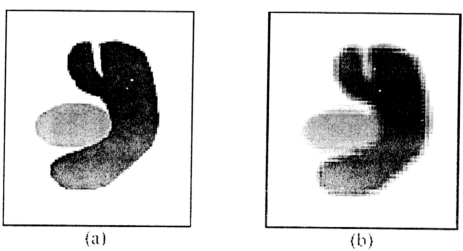


Figure 2.14 - Partial Volume Effect (a) Ideal Image (b) Acquired Image.

Segmentation methods can be divided into three categories depending on the level of automation. The variations of these types of interactions are the amount of time, effort and the training required. Manual methods are vulnerable to reliability issues. However, sometimes, even "automated" segmentation require specification of initial

parameters that effect the performance. Segmentation methods include :-

1. ***Manual*** – Manual segmentation is done by hand using aprior knowledge and human visual system to make qualitative analysis of images in order to recognize various anatomical structures and tissues. As mentioned, manual segmentation requires knowledge of the domain (aprior knowledge) [40,41], is time consuming to segment image by image [42,28], costly [43,28] and a lot of effort is required. As the manual segmentation results depend on the individual doing the segmentation, it leads to poor reproducibility of the results [42,43].

2. ***Semi-Automatic*** (interactive) – Semi-automatic segmentation requires human input in the process. The process is automatized after the key data is known. The combination of human knowledge and computing capabilities makes it possible to process large data sets [44]. Users can point out the anatomy of interest, and the computer can perform the segmentation based on the information given [44,45,29].

3. ***Automatic*** – Automatic segmentation results are highly reliable, consistent and reproducible. Automatic segmentation methods focus on very specific tasks and data, while manual and semi-automatic segmentation are more flexible to segment various anatomies. In [38], the segmentation algorithm is specific to segment brain tissue, while in [39,40,46], the anatomy to segment it selected by the user [47]. Wells *et. al.* in [48] used Expectation-Maximization (EM) algorithm to automatically segment healthy brain tissues. Kshirsagar *et. al.* in [42] gave a fully

automated technique for obtaining measurements of the femoral cartilage thickness of the human knee joint.

Segmentation is divided into three major classes by [2] based on the information manipulated in the images-

1. ***Point Based Segmentation*** - Point based segmentation methods classify a voxel depending only on its intensity, no matter where it is located. Thresholding is a simple example of this method. Pattern recognition methods have successfully developed to segment MRI data of head and chest. But there are instances where isolated voxels or small regions are incorrectly classified.

2. ***Edge Based Segmentation*** - Edge based segmentation methods detect intensity discontinuity in gray level volume. The edges are assumed to represent the border between different organs or tissues, and the enclosed areas are defined as regions. There are 2 types of edge based approaches [49]:-
   - Density based approach that look for isodensity contours or surfaces
   - Gradient based approach that looks for large changes in the gradient.

The common edge based method is to locate the maxima of the first derivative of the intensity function. However, contours detected this way usually are not closed. Examples include Laplacian, Prewitt, Sobel, and Roberts gradient based edge detectors [49,34,50,33]. Another option is to use zero crossing of second derivative to detect edges. These operators sometimes create erroneous bridges between

different materials that have to be removed interactively. Examples are the Marr -Hildreth [132] and Laplacian of the Gaussian operators [2].

3. ***Region Based Segmentation*** - Region based segmentation methods consider whole regions instead of individual voxels or contours. Factors such as size, shape, location, variance of gray levels and spatial relations to other regions are taken into consideration. Typically region based methods post-process the results of point based segmentation. This step is to reduce misclassification of voxels in smaller regions. Region growing algorithms can also be used to split and merge regions [51]. Segmentation and interpretation steps are combined into one region base segmentation step. Human users or models can be used to provide the knowledge required for segmentation [44,52,53]. Automatic region based segmentation would give incorrect answer if the underlying model do not represent the data properly. Thus far, biological models are not adequate to handle various pathologies.

2.5.1 MAGNETIC RESONANCE IMAGING (MRI) SEGMENTATION

MRI has the ability to derive contrast from a number of tissue parameters that there are many pulse sequences for acquiring MR images. It is important to determine the optimal pulse sequences to obtain accurate segmentation. This problem requires the knowledge of the underlying tissue properties of the anatomy to be segmented. Different subjects may require different pulse sequences because of the anatomical and physiological variability. Simmons *et al.* in [54] gives the following factors to take into consideration for successful segmentation:-

- Choosing the appropriate spin sequence to enhance tissue contrast that is being segmented.
- Preprocessing images to compensate for non-uniformities, noise, correction bias and aliasing.
- Choosing the appropriate segmentation process.

Segmentation of MR images are divided into the following categories in [28,17] :

1. Thresholding approaches
2. Edge-based segmentation approaches
3. Region growing approaches
4. Classifiers
5. Clustering approaches
6. Markov random field (MRF) model
7. Artificial neural network (ANN)

8. Deformable models

9. Atlas guided approaches

10.Hybrid approaches

2.5.1.1 THRESHOLDING

Thresholding segments scalar images by creating a binary partitioning of the image intensities [28]. Figure 2.15 shows the histogram of a scalar image that possesses three classes corresponding to three models. A threshold procedure attempts to determine the threshold value, which separates the desired classes. Segmentation is achieved by grouping all pixels with intensity smaller that the threshold into one class, and all the other pixels into another class [28]. Figure 2.15 shows two potential thresholds at the valleys of the histogram. Multi-thresholding segments with more that one threshold value [28].

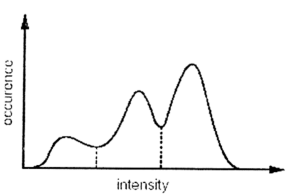


Figure 2.15 - A Histogram Showing Three Apparent Classes

Thresholding is an effective method to segment images where different structures have contrasting intensities or other quantifiable features. The threshold values are

usually generated interactively, through automated methods do exist [17,28]. Interactive methods can be based on an operator's visual assessment of the resulting segmentation when the thresholding is implemented in real-time.

Thresholding is the primary step in a sequence of image processing operations. Thresholding is limited, reason being it generates only two classes in its simplest form. Furthermore, it cannot be applied to multi-channel images [28]. It also does not take into consideration the spatial characteristics of an image, which causes it to be sensitive to noise and inhomogeneities in MR images. These artifacts corrupt the histogram of the image, making separation difficult. Variations of classical thresholding have been proposed that incorporate information based on local intensities and connectivity [28].

Global thresholding has been performed on CT data. Local thresholding has been applied to MRI data in combination with morphological filtering [40].

2.5.1.2 EDGE-BASED APPROACHES

Edge based schemes experience incorrect detection of edges due to the noise, over- and under segmentation, and variability in threshold selection in the edge image[17]. Bomans *et al.* in [132] combines Marr-Hildreth operator with morphological filtering, which requires manual labeling and editing for 3D display. Another method is boundary tracing. This method is restricted to segmentation of large, well defined structures but not individual tissue types [17]. Kshirsagar *et al.* in [42], uses

Canny's criteria for edge detection and template matching on MR images of human the knee joint to measure femoral cartilage thickness.

2.5.1.3 REGION GROWING APPROACH

Region growing is a technique for extracting a region of the image that is connected on some predefined criteria, which can be based on intensity information and/or edges of the image [28]. Region growing requires a seed point that is manually selected. It extracts all pixels connected to the initial seed with the same intensity value (Figure 2.16) [28].

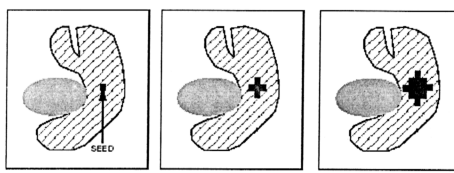


Figure 2.16 - An Example of Region Growing

Region growing is often used within a set of image preprocessing operations for delineation of small, simple structures such as tumors and lesions [28]. The disadvantage is that it requires manual interaction to obtain the seed point for each region to be extracted. Split and merge algorithms are related to region growing but do not require seed points [28]. Region growing also can be sensitive to noise, causing extracted regions to have holes or to be disconnected [28]. Alternatively,

partial volume problems can cause separate regions to become connected. A homotopic region growing algorithm has been proposed to preserve the topology between the initial region and the extracted region.

Commercial software like ANALYZE by Mayo Foundation [58], and MIMICS by Materialise [59] use seed growing algorithm for segmentation and 3D reconstruction. Results obtained with seed growing algorithm depend on the operator settings. Sethian in [133], used level set method to extract anatomical features from medical images. The algorithm propagates from the initial seed in the medical image. When the interface passes over places where the image gradient is small, it is assumed not a boundary. Where the image gradient is large, the algorithm suspects a boundary and slows the curve down. The fast matching version of level set is used to obtain optimal performance.

2.5.1.4 CLASSIFIERS

Classifier methods are pattern recognition techniques that seek to partition a feature space derived from the image using data with known labels [28]. A feature space is the range space of any function of the image, such as the image intensity or texture [28]. Histogram is an example of a 1-D feature class.

Classifiers are known as supervised methods because they require training data that are manually segmented and used as references to automatically segmenting new data [28]. Nearest-neighbor classifier is a simple classifier, where each pixel or voxel is

classified in the same class as the training datum with the closest intensity. The *k*-nearest-neighbor (kNN) classifier is a generalization of this approach, where a pixel is classified according to the majority vote of the *k* closest training data. kNN does not make any underlying assumptions about the statistical structure of the data, so it is considered as a non-parametric classifier [17,28]. Manduca *et al.* in [46] developed tools for MRI data segmentation using kNN.

Maximum likelihood (ML) or Bayes theorem is a parametric classifier. It assumes that the pixel intensities are independent samples from a mixture of Gaussian probability distributions. A pixel is assigned to a class with the nearest posterior probability. The ML classifier performs well when the data follows a Gaussian distribution. It is capable of providing a soft segmentation composed of the posterior probability [28]. Bayesian classification is used in [37] to classify material mixtures in MR volume data using voxel histogram.

Classifiers require the structure to be segmented posses distinct quantifiable features. Since, training data can be labeled, classifiers can transfer these labels to the new data. They are computationally efficient and can be applied to multi-channel images [28]. Classifiers do not perform spatial modeling. Moreover it also requires manual interaction to obtain training data, which can be time consuming and laborious. The use of the same training set for a large number of scans can lead to biased result, that does not take into account anatomical and physiological variability between subjects.

2.5.1.5 CLUSTERING APPROACH

Clustering algorithms are unsupervised methods as they perform the same function as classifier without the use of training data. Clustering methods iterate between segmenting the image and characterizing the properties of each class. They train themselves using the available data.

Common clustering algorithms are the K-means algorithm, fuzzy C-means algorithm and the expectation-maximization (EM) algorithm [28]. K-means algorithm clusters data by iteratively computing a mean intensity for each class and segmenting the image by classifying each pixel in the class with the closest mean [28]. Fuzzy C-means algorithm generalizes K-means algorithm, by allowing for soft segmentation based on fuzzy set theory [43]. EM algorithm assumes that the underlying data follows a Gaussian mixture model. It iterates between computing the posterior probabilities and computing maximum likelihood estimation of the means, covariances and mixing coefficients of the mixture model.

Clustering algorithms do require initial parameters. They do not incorporate spatial information, so they are sensitive to noise and intensity inhomogeneities. This allows for fast computation [28].

Some of the research done using clustering algorithms are as following :-

- Dzung Pham *et al.* in [43] performed fuzzy C-means clustering to obtain a spatial map representing the membership values of the different tissues present in the MR

brain image.

- Wells *et al.* in [47] and [48] used expectation-maximization (EM) algorithm to estimate tissue classes and to correct gain field. This method is fully automated for segmenting healthy brain tissue. For a given type of acquisition, intensity variation across patients, scans and equipments have been accommodated without manual intervention in the segmentation algorithm. The segmentation was compared to manual segmentation and supervised multi-variate classification segmentation. Wells *et al.* [47,48] reported that this method was found to be consistent with manual segmentation, and closer to manual segmentation than supervised methods.
- Simmons *et al.* in [54] and Manduca *et al.* in [46] also reported the use of clustering approach to segment MR images.
- Hall *et al.* in [60] used adaptive fuzzy rules to identify voxels from MR images before clustering. This allowed clustering to be done on a subset of an image with initialization. The identified voxels can also be used to identify clusters.
- Clark *et al.* in [61], Fletcher-Heath *et al.* in [62] and Clarke *et al.* in [63] used unsupervised fuzzy clustering and domain knowledge to segment brain tumors in MR images.

2.5.1.6 MARKOV RANDOM FIELD (MRF) MODELS

Markov random field (MRF) modeling is not a segmentation method but a statistical method used within the segmentation method [28]. MRF models spatial interactions between neighboring or nearby pixels, which provide a mechanism for modeling a variety of image properties. MRF is used by taking into account that most pixels

belong to the same class as their neighbors. MRF is incorporated into clustering algorithm such as k-means algorithm under Bayesian prior model.

MRF models have difficulty selecting proper parameters to control the strength of spatial interactions [28]. High setting can result in an excessively smooth segmentation and loss of important details. MRF is also used to model intensity inhomogeneities that occur in MR images and texture properties [28].

Zhang *et al.* in [64] proposed a novel hidden Markov random field (HMRF) model for segmentation of MR brain images. HMRF is a stochastic process generated by a MRF models whose state sequence cannot be observed directly but which can be observed through observation. An EM algorithm was used to fit the HMRF model.

2.5.1.7 ARTIFICIAL NEURAL NETWORK (ANN)

Artificial neural networks are massively parallel networks of processing elements or nodes that simulate biological learning [28], where each node is capable of performing elementary computation. Adaptation of weights assigned to the connections between nodes facilitates learning.

ANN can be used in a variety of ways to segment images. They can be used as classifiers, where the weights are determined using training data, as unsupervised clustering methods and as deformable models [28].

Spatial information can be incorporated into the ANN classification procedure due to its many interconnections. Even though ANNs are parallel, but they are simulated on standard computers, thus reducing computational advantages.

Some of the work done using ANN are :-

- Segmentation of MR image using 3 layer feed forward ANN and probabilistic ANN by Manduca *et at.* in [46].
- Segmentation of medical images using Competitive Hopfield neural network (CHNN) imposed by winner-takes-all learning mechanism, based upon global information of the gray levels distribution by Cheng *et al.* in [65]. CHNN method shows promising results in comparison with fuzzy c-means method.
- Segmentation of MR brain images using Hopfield ANN by Sammouda *et al.* in [66]. Sammouda *et al.* showed that the Hopfield ANN has clear advantage when compared to the Boltzman Machine and ISODATA clustering technique.
- Segmentation of MR images of the brain using Learning Vector Quantization (LVQ) ANN by Alirezaie *et al.* in [67]. The results were compared with Back-Propagation ANN. LVQ ANN is found to be insensitive to gray level variation of MR images between different slices while performing faster and better than Back-propagation ANN.
- Segmentation of MR images using probabilistic neural network by Toulson *et al.* in [68].
- Segmentation of MR images of the femur using wavelet and self organizing maps (SOM) ANN by Woo *et al.* in [69] and Woo in [20].
- Segmentation of multi-spectral medical images using fuzzy ISODATA by

McMahon *et al.* in [39].

- Automatic segmentation and tissue classification of anatomical objects from MRI data sets using Kohonen ANN by Busch *et al.* in [70] .
- Segmentation of MR brain image using pulsed-coupled neural network (PCNN) by Keller *et al.* in [71]. PCNN incorporates both spacial and intensity values unto the thresholding process. PCCN does well at contrast enhancement but requires a lot of manual intervention to produce the desired results. PCNN does well in image segmentation only when each segment is approximately uniform in intensity. PCNN requires the proper setting of various parameters so that a uniform respond is achieved over a set of images.

2.5.1.8 DEFORMABLE MODELS

Deformable models are physically motivated, model-based techniques for delineating region boundaries using closed parametric curves or surfaces that deform under the influence of internal and external forces [28]. A closed curve or surface is placed near the desired boundary and is allowed to undergo an iterative relaxation process. Internal forces are computed from within the curve or surface to keep it smooth throughout the deformation. External forces are derived from the image to drive the curve or surface towards the desired feature of interest [28]. Figure 2.17 shows an example of applying a 2-D deformable model or active contour to a MR heart image. The active contour was initialized as a circle and allowed to deform to the inner boundary of the left ventricle. A deformable model moves according to its dynamic equations and seeks the minimum of an energy function [28].

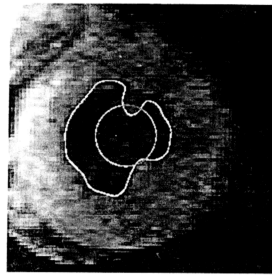


Figure 2.17 – Deformable Contour - A deformable contour is used to extract the inner wall of the left ventricle of a human heart from an MR image. The initial deformable contour is plotted in gray and the final converged result is plotted in white.

Deformable models have the ability to directly generate closed parametric curves or surfaces from images and they incorporate a smoothness constraint that provides robustness to noise and spurious edges [28]. However, they require manual interaction to place an initial model and choose appropriate parameters. Standard deformable models also exhibit poor convergence to concave boundaries [28].

Some of the research done using deformable models are :-

- Safont *et al.* in [72] used active contours to reconstruct the proximal femur from CT scans. Active contours were used, as classical methods failed because there was no contrast between the different bones.

- The active contour model in [73] by Durikovic *et al.* uses the texture information to segment small areas of tissue.
- Behiels *et al.* in [74] used active shape models (ASM) to segment bone structures in radiographs images.
- Cootes *et al.* in [75] also used ASM for locating structures in medical images. A compact model with the shape and appearance of flexible organs seen in 2D images are built. These models are parameterised for configuration.
- Ghebreab *et al.* in [76] used a string technique to segment the medical image of the human lumbar. The string is a variational deformable model, learned from the collection of example objects. This work is extended in [77], where landmarks are used to segment spine structures. The landmarks are integrated into the well-known deformable models.
- Rifai *et al.* in [78] used deformable models to segment the bone region in MRI brain volume data set, taking into account the partial volume effect.
- Székely *et al.* in [79] used snakes to segment 2-D and 3D objects from MRI volume data.
- A framework for automatic model extraction from MR images was proposed by Lötjönen *et al.* in [80]. The frame work has two stage :- the construction of multi-resolution model using pyramid graphs and the matching the model with the input data.
- Lou *et at.* in [81] proposed an algorithm for the semi-automatic segmentation of medical image series based on the combination of the live wire algorithm and the active contour model.
- Alphonso in [82], used ASM models to segment the valves in the human heart

using MRI images. The segmented images were then used to create a 3D image to depict the valve movement in the human heart.

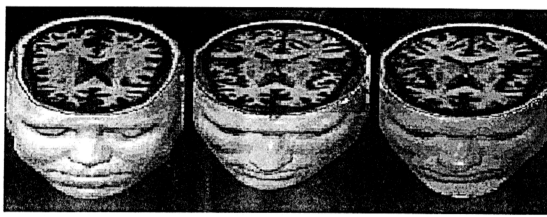
- Alexandre *et al.* in [55] presents two boundary based paradigms called live wire and live lane for medical image segmentation. In live wire approach, the user initially specifies a point on the boundary using the cursor. For any subsequent position of the cursor, a curve connecting the initial point and the current cursor point is calculated. As the cursor closes the boundary, the live wire snaps onto the boundary. In the live lane approach, the user involvement is more active and tightly integrated with the machine's action. To extract a boundary in a given slice, the user only selects an initial point and subsequently steers the cursor in the vicinity of the boundary within a lane of certain width. Alexandre *et al.* in [56], extends the live-wire method to 3D image segmentation. An ultra fast life wire method was introduced in [57] by Alexandre *et al.*, which is 1.3-3.1 times faster that live wire.

2.5.1.9 ATLAS GUIDED APPROACH

Atlas-guided approaches are powerful tools for medical image segmentation when a standard atlas is available [28]. The atlas is generated by compiling information on the anatomy of interest, and then the atlas is used as a reference frame for segmenting new images. Atlas-guided approaches are similar to classifiers except they are implemented in the spatial domain rather than in a feature space.

Standard atlas-guided approaches treat segmentation as a registration problem. Atlas warping finds a one-one transformation that maps a pre-segmented atlas image to the

target image that requires segmenting. Because of anatomical variability, a sequential application of linear and non-linear transformation is used [28]. Figure 2.18 shows an example of atlas warping for MR head scan. All structural information is transferred to the target image because the atlas is already segmented.



(a) (b) (c)

Figure 2.18 - Demonstration of Atlas Warping: (a) template image, (b) target image (c) warped template [28]

Atlas-guided approaches have been applied mainly to MR brain images [28]. Atlas-guided transfers labels as well as segmentation, while providing a standard system to study the morphometric properties [28]. Accurate segmentation of complex structures is difficult even with non-linear registration due to anatomical variability. So, atlas-guided approaches are generally better suited for segmenting structures that are stable over the population of study [28].

Pien *et al.* in [83] used MR image of the knee to determine the volume thickness of cartilage of the knee with minimal operator intervention. A 3D digital template of

normal knee anatomy is segmented by hand. The template is then used for segmentation. Then cartilage thickness is determined by computing the smallest distance between the surfaces of the cartilage at every surface voxel.

2.5.1.10 HYBRID APPROACHES

1. Kapur *et al.* in [84] used atlas guided approach and Bayesian classifier to segment MRI data of the knee.
2. Lin *et al.* in [85] carried out segmentation of MR brain images using fuzzy clustering and Hopfield neural network. The proposed Fuzzy Hopfield neural network (FHNN) results are more promises that hard c-mean method, especially where artifacts can be resolved using FHNN. FHNN differs from the conventional Hopfield NN in that a fuzzy c-mean clustering strategy is imposed for updating the neuron states.
3. Kapur in [38] used EM segmentation, binary morphology and active contours models to segment brain tissue from MR images.
4. Blonda *et al.* in [86] used fuzzy Kohonen clustering network (FKCN) to segment multi-spectral MR brain images. Then a feed forward ANN based on back propagation learning rule is used for tissue labeling.
5. Germond *et al.* in [87] proposed the use of multi-agent system, a deformable model and an edge detector to automatically segment MRI brain images.
6. Pemmaraju *et al.* in [88] used multi-resolution wavelet decomposition to reconstruct the original image such that it contains all the salient features relevant to segmentation and is devoid of the low frequency noise and texture information that can be ignored while segmenting the image. An unsupervised neural network

with fuzzy learning rules is then used to segment the reconstructed image.

2.5.1.11 OTHER APPROACHES

1. ***Morphological Operator*** – Hohne *et al.* in [44] and Schiemann *et al.* in [45] used low-level morphological operators to interactively segment MRI and CT volumes. They implemented simple binary operations such as thresholding, connected component analysis, dilation, erosion, region filling and Boolean operation for interactive segmentation. This work was then extended to the segmentation of the male Visible Human [89].

2. ***Model Fitting*** – Model fitting is a segmentation method that fits a simple geometric shape to the location of extracted image feature [28]. It is specialized to the structure being segmented, but it is easily implemented and provides good results when used appropriately [28]. A general approach is to fit splines, curves or surfaces to the feature. However, with model fitting, the image feature must first be extracted before the fitting can take place.

3. ***Watershed*** – Watershed algorithm uses concepts from mathematical morphology to partition the image into homogeneous regions [28]. Over segmentation occurs when the image is segmented into unnecessarily large number of regions. Thus, watershed algorithm in medical imaging is usually followed by a post-processing step to merge separate regions that belong to the same structure [28]. Nguyen *et al.* in [90] showed how to represent watershed segmentation as an energy

minimization problem, using the distance-based definition of the watershed line. Apriori consideration about smoothness is then imposed by adding the contour length to the energy function. This new method is called watersnakes.

4. ***Landmarks*** – Ghebreab *et al.* in [91] uses landmarks to segment volumetric vertebrae from CT images by reducing segmentation complexity. Point landmarks, curve landmarks and sheet points are differentiated based on the number of dimensions in which boundary points are well defined. These are input to a priority segmentation scheme which tries to find complete boundaries in the image. A deformable model is used to implement this method. Bokde *et al.* in [131] also used landmark based segmentation of MR brain images.

2.5.2 SEGMENTATION VALIDATION

Validation experiments are necessary to quantify the performance of a segmentation method. There are five methods that can be used for validation [17,28]:-

1. Correlation of automating segmentation with histopathology studies.
2. Comparing the automated segmentation with the use of physical phantoms
3. Comparing the automated segmentation with the use of computational phantoms.
4. Reproductively studies
5. Manual labeling

Histopathology – This techniques, that only compares volume, gives a good measure of accuracy but is not adequate for validation of segmentation since it do not confirm the location or the shape of the anatomy of interest. Histopathology also has some morphometric metric difficulties [17]:-

- Pre-mortem data does not correspond well with pathology because of logistics of the excision
- MR relaxation behavior of excised tissue is very different from perfused anatomy
- The work is very labor intensive.

Pathology correlation may not be the feasible way for verification of segmentation methods, even with their apparent value for ground truth.

Physical phantoms containing paramagnetic liquid and gel doped with paramagnetic agents have been introduced to mimic MRI parameters of the tissues being modeled

[17]. Physical phantoms provide an accurate depiction of the image acquisition process, but do not allow realistic segmentation due to high level of geometric complexities in 3D and multiple classes present in human anatomy [17,28]. Phantoms do not exhibit the characteristics that make segmentation of human tissues so difficult. Table 2.4 reports the accuracy levels obtained in various physical phantoms and other verification studies [28]. Phantoms display a wide range of accuracy depending on the method of choice [17,28]. Although phantoms images provide excellent means for daily quality control of the MR scanner, but they provide only a limited degree of confidence in the reliability of the segmentation methods. Phantom measurement cannot realistically be used to clinically validate MRI segmentation, but should be used to rule out invalid segmentation rising from variability in MRI system performance [17].

Computational phantoms are more realistic, but only simulate the image acquisition process using simplified models [28]. Computational phantoms are also extended to investigate a variety of MR processes such as optimization of RF pulse technique [17]. The robustness of the segmentation process may be probed by corrupting the simulated signal with noise, nonlinear field gradient and non-uniform RF excitation [17]. This way one source of uncertainty can be introduced and the resulting segmentation can be related to the uncertainty in a quantifiable manner. Clarke *et. al.* in [17] believed that computer generated phantoms provide a good way to quantify segmentation methods.

Table 2.4 – Comparison of Accuracy Reported for Various Phantom and Other Verification Studies [28]

Author	Year	Method	Type	Accuracy (%)
Jack	1991	Manual tracing / thresholding	Phantoms	3
Ashtar	1990	Operator guided boundary tracing	Phantoms	2
Kohn	1991	Linear decision boundary in 2D feature map	Phantoms	9
Cline	1991	KNN	Phantoms	7
Gerig	1992	Maximum likelihood	Phantoms	3
Peck	1992	Eigen image	Phantoms	2
			Egg Yolk	12
Rusin	1993	Eigen image	Phantoms	5
Jackson	1993	kNN	Phantoms	2
Vinitski	1994	kNN	Phantoms	9
Mitchell	1994	kNN or maximum likelihood	Phantoms	10
Snell	1994	Active template matching	Dissected brain	6
			Cadaver	10

Reproducibility has been measured under different variations [17]: -

- Usage of the same data, however having the same operator select several training data sets (intra-observer variations)
- Using different operators to perform the segmentation (inter-observer variation)
- Stability for multiple scans of the same subject using different imaging sessions
- Variation of the segmentation over different patients

Reproducibility does not provide a measure of accuracy, but simply gives a measure of reliability under the variations that is tested. The reproducibility of normal tissue is dependent on the segmentation method, targeted time and the size of the tissue volume Reproducibility of reported lesion volume depends on lesion type, staging and size [17].

Manual Labeling uses experts to manually trace the boundaries of the different tissue regions. It truly mimics the radiologist interpretation, yet it does not guarantee a perfect truth model since an operator's performance can also be flawed [28,17]. It is also labor intensive and time consuming for large data sets.

2.5.3 HUMAN FEMUR

The human skeleton consists of 206 bones that are bound together by ligaments. Figure 2.19 illustrates the human skeleton. A living bone consist of three layer:- the outside skin, the hard compact bone and the bone marrow [5]. The male and female skeletons are fundamentally the same. Female bones are generally lighter and thinner that male bones. Female pelvis is shallower and wider than the male pelvis for childbirth purposes.

Figure 2.19 – Human Skeleton

Femur is also known as the thighbone or upper bone of the leg or hindleg. It is the longest bone is the human skeleton [6,5]. . Figure 2.19 shows the location of femur in the human skeleton. Figure 2.20 shows the different parts of the human femur. The femur head forms a ball-and-socket join with the hip (at the accetabulum) [6,5]. The neck of the femur connects the shaft and the head at a 125^0 angle, for efficient walking [5,6]. Two large prominences, or condyles, on either side of the lower end of the femur form the upper half of the knee joint [5,6]. Human femur has shown to be capable of resisting compression forces of 1,800 – 2,500 pounds [6]. It is stronger than reinforced concrete [5].

Diseases and injuries of bone are major cause of abnormalities of the locomotor system. The femur is subjected to injuries, like any other bone in the body. Examples of injuries are dislocation and fracture [5]. Diseases such as arthritis and osteoporosis can also damage the femur. These injuries and diseases are very agonizing and/or cause nonfunctional joint movements.

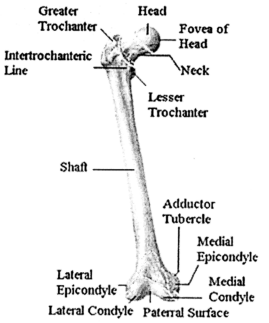


Figure 2.20 – Human Femur

2.5.4 MAGNETIC RESONANCE (MR) BONE IMAGE SEGMENTATION

It is difficult to segment MR images [83,48,68,47,84,69,67] because of the number of tissue contrast. In some cases, different tissues have the same range of intensities or the same tissues have different intensities. Figure 2.21 gives an example of MR image of the femur with with a big range of pixel values in a single slice. Segmentation of MR femur image is further complicated because femur is an organic structure that is asymmmetric.

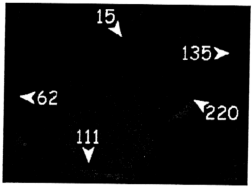


Figure 2.21 - Example of a MR Image of the Femur with Varying Bone Pixel Values

Alirezaie *et. al.* in [67] gave the different variations present in medical images :-

- Intra-slice variation
- Inter-slice variation
- Inter-patient variation

There are countless papers written on MRI segmentation of the brain [98]. However, there are only limited number of papers written on bone segmentation using MR images. Some the previous work done on MR femur image segmentation are discussed below :-

- Woo in [20] uses a hybrid algorithm of wavelet transform (WT) and SOM neural network to segment MR femur transverse images. Multi-resolution WT is initially applied on the region of interest (ROI) as to reduce the dimensionality of the ROI. Femur extraction from the lowpass band of the resultant image is done using a SOM neural network. In order to enrich the set of pixels constituting the bone

extracted from the lowpass band, a statistical classification of the residual high frequency bands belonging to the isolated image was performed. The distributed pixels scattered in the high frequency bands, belonging to fragments of the femur were compiled through statistical classification. This technique provided encouraging results for images with "thick" bone regions. However, "thin" bone regions produced partial bone extraction.

- Nagappan in [19] used two different strategies to segment the femur. An automatic segmentation algorithm was used for MRI slices where the femur bone is "thick" and the quality of the images were good. Manual segmentation is performed on MRI slices that contain "thin" femur bone and are of poor quality. The automatic segmentation used is based on wavelet transforms and SOM neural network [20]. The segmented slices are stacked up into a volume image after interpolation. The volume is then put through a modified marching cubes algorithm to extract out the surface information of the femur. Finally, this surface information was pre-processed before being sent to a rapid prototype system for the femur to be fabricated.

- Mangalam *et al.* in [92] concentrated on identification and classification of different materials in MRI volume data of the femur. The MRI slices were converted from 2D images to a volumetric 3D image. A histogram of the volumetric image was defined. Curve fitting using splines was performed on the histogram. Spline fitting is required to reduce the effects of noise from the histogram. It helps to smoothen the histogram for Gaussian fitting. Each material

present in the image was represented as a Gaussian function. Bayesian theorem was then applied to classify materials. The goal of Bayesian theorem was to classify the materials in such a way to minimize the probability of misclassification [130].

- Raveendran *et al.* in [93] used Sobel, Roberts and Delta operators to segment the bone from MR femur image.

2.5.5 APPLICATIONS OF MEDICAL IMAGE SEGMENTATION

Diagnostic imaging is not the only major field of application for segmentation. It is usually a pre-processing step for other medical applications such as surgery planning, diagnostic, anatomical studies or biomodelling [37]. 3D segmentation and imaging is required in fields where therapeutical decisions have to be made by non-radiologists on the basis of radiological images. Here are some of the fields where segmentation and imaging can be used :-

1. ***Diagnostics*** – MR image segmentation is used in the diagnosis of multiple sclerosis, tumors of the pituitary gland and brain, infections in the brain, spine or joints, tendonitis and stroke in the earliest stages [25].

2. ***Surgery Planning*** - Craniofacial surgery is a major application of 3D imaging. MIMICS software by Materialise [59] is a software used extensively in Europe for craniofacial surgery. Neurosurgery planning is an application that is becoming

more attractive with the use 3D MR images [94]. 3D visualization of brain tumors allow the surgeon to find a minimal risk path before surgery [2,8].

3. ***Traumatology*** - In emergency situations, planning time is very short. 3D imaging techniques are introduced in difficult cases as new faster imaging modalities are available and computing power increased [2]. For example, in pelvis surgery, 3D imaging is considered an enormous help to access difficult morphologies.

4. ***Radiotherapy Planning*** – In radiotherapy planning, the objective is to focus the radiation to the target volume and avoid healthy organs. A realistic rehearsal of the treatment procedure allows the medical examiner to visualization the target volume. With segmentation methods, medical personal would be able to measure the target volume and have a quantitative record of the effectiveness of the treatment.

5. ***Implant Design*** – 3D simulation of interaction between an implant and the bone helps the surgeon evaluate the implant effectiveness. Finite element stress analysis can be performed to predict the long term outcome of implants [8]. Medical device manufacturers through the use computer aided design (CAD), design accurate implants that conform to the shape of the body. Medical devices such as hearing aids, surgical tools and orthopedic implants benefit from the use of anatomical models [95]. Bio-models also help surgeons plan for surgery. Custom made implants and prothesis for a patient reduces recovery and surgery time,

while in the long term helps to reduce cost [19,8]. Product engineers on the other hand can see how the device fit in the body early in the design process, improving ergonomics and reducing product redesigning steps [95].

6. ***Quantification of Tissue Volumes*** – MR images give good soft tissue contrast. This property is used in tumor diagnostics. This step is extended further with 3D volume imaging. Researchers are able to segment tumor tissue from healthy tissues [96]. The quantity or the percentage of tumor tissue helps researchers to plan treatment for patients. Rescanning allow researcher to have a quantitative measurement of the tumor tissue and its respond to the treatment given.

7. ***Robotic Surgery*** – Segmentation of human anatomy with visualization tools are used in robotic surgery. Surgeons are able to use remotely controlled "microbots" – milimeter sized tethered robots - to probe into the patient's body to remove cancer cells [135].

8. ***Bio-modeling*** – Another application of segmentation is bio-modeling. Medical images are used to model the anatomy of interest. The anatomy is segmented first, and the 3D model is constructed in the computer. The bio-models are then made using rapid prototype technology [139].

9. ***Medical Research and Education*** - Apart from clinical work, 3D imaging is used in medical research and education. Voxel Man [97] is a multimedia software that

is used in anatomical and radiological studies. The image for this program is derived from the cross-sectional images of the Visible Human Project. This program allows interactive exploration of detailed 3D anatomical models. It presents the radiological manifestation of the normal anatomy.

2.6 ARTIFICIAL NEURAL NETWORK (ANN)

Figure 2.22 illustrates the flow of Section 2.6. The definition of ANN is given in Section 2.2.1. The history of ANN is explained briefly in Section 2.6.2. The benefits of ANN are discussed in Section 2.6.3 while the difference between ANN and traditional computing are specified in Section 2.6.4. The analog of the brain with ANN is clarified in Section 2.6.5. Neurons, the building block of ANN, are construed in detail in Section 2.6.6. The different architectures of ANN are illustrated in Section 2.6.7. The different methods to train an ANN are presented in section 2.6.8. The uses of ANN are addressed in Section 2.6.9. Finally, self organizing maps (SOM) ANN is explained in detail in Section 2.6.10.

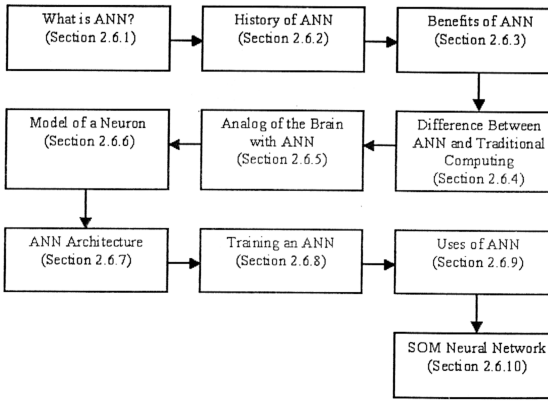


Figure 2.22 – Overview of Section 2.6

2.6.1 WHAT IS AN ARTIFICIAL NEURAL NETWORK (ANN)?

Artificial neural networks (ANNs), commonly referred to as "neural networks" (NN), are electronic models based on the neural structure of the brain [104]. It has been motivated by the recognition that the human brain computes entire differently from the conventional computer. The brain is a highly complex, nonlinear, parallel computer. The brain organizes neurons to perform computation faster than digital computers. A NN is a machine that is designed to simulate the way the brain performs a particular task or function of interest. A neural network can be defined as a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use [100,101,102,103]. It resembles the brain in two respects [103]:

1. Knowledge is acquired by the network from its environment through a learning process.
2. Inter-neuron connection strength, know as synaptic weight, is used to store the acquired knowledge.

2.6.2 HISTORY OF ARTIFICIAL NEURAL NETWORKS (ANN)

The first step toward artificial neural networks came in 1943 when Warren McCulloch, a neurophysiologist, and Walter Pitts, a young mathematician, wrote a paper on how neurons might work [105]. They modeled a simple neural network with electrical circuits. Donald Hebb wrote the Organization of Behavior in 1949 to reinforce this concept of neurons and how they work [104]. The paper pointed out that neural pathways are strengthened each time that they are used [102].

As computers advanced into their infancy of the 1950s, it became possible to begin to model the rudiments of these theories concerning human thought. Nathanial Rochester from the IBM research laboratories led the first effort to simulate a neural network [104], which failed. However later attempts were successful. It was during this time that traditional computing began to flourish and, as it did, the emphasis in computing left the neural research in the background.

Throughout that time, advocates of "thinking machines" continued to argue their cases. In 1956 the Dartmouth Summer Research Project on Artificial Intelligence (AI) provided a boost to both artificial intelligence and neural networks [104]. This project stimulated research in the AI field and in the neural processing part of the brain.

In the years following the Dartmouth Project, John von Neumann suggested imitating simple neuron functions by using vacuum tubes [104]. Frank Rosenblatt, a neuro-

biologist of Cornell, began work on the Perceptron [104]. He was intrigued with the operation of the eye of a fly, where much of the processing for the fly to flee is done in its eye. The Perceptron, which resulted from this research, was built in hardware and is the oldest neural network currently in use today. The Perceptron computes a weighted sum of the inputs, subtracts a threshold, and passes one of two possible values out as the result. The perceptron's limitations were proven by Marvin Minsky and Seymour Papert's 1969 book *Perceptrons*.

In 1959, Bernard Widrow and Marcian Hoff of Stanford developed models they called ADALINE (ADAptive LINear Elements) [106,102] and MADALINE (Multiple ADAptive LINear Elements) [104]. MADALINE was the first neural network to be applied to a real world problem. It is an adaptive filter which eliminates echoes on phone lines. This neural network is still in commercial use.

These earlier successes allowed people to exaggerate the potentials of neural networks. This excessive hype infected the general literature of the time. Disappointment set in as promises were unfilled. Writers also began to ponder what effect "thinking machines" would have on man. Isaac Asimov's fictional series on robots revealed the effects on man's morals and values when machines where capable of doing all of mankind's work [104]. These fears, combined with unfulfilled claims, caused respected voices to critique the neural networks research, causing to halt of the funding. This period of stunted growth lasted until 1981.

In 1982, John Hopfield of Caltech presented a paper to the national Academy of

Sciences [101,104]. Hopfield's approach was not to simply the model of the brain but to create useful devices. He showed how such networks could work and what they could do with clarity and mathematical analysis.

At the same time, a US-Japan Joint Conference on Cooperative/Competitive Neural Networks was held in Kyoto, Japan [104]. Japan subsequently announced their Fifth Generation effort. US periodicals picked up that story which generated a worry that the US could be left behind causing funding to flow once again.

By 1985 the American Institute of Physics began what became an annual meeting - Neural Networks for Computing. By 1987, the Institute of Electrical and Electronic Engineer's (IEEE) first International Conference on Neural Networks drew more than 1,800 attendees [104].

The 1990 US Department of Defense Small Business Innovation Research Program named 16 topics which specifically targeted neural networks with an additional 13 mentioning the possible use of neural networks [104].

Today, neural networks discussions are occurring everywhere. Currently most of neural networks development is simply proving that the principal works. Companies are working on three types of neuro chips - digital, analog, and optical. Some companies are working on creating a "silicon compiler" to generate a neural network Application Specific Integrated Circuit (ASIC) [104]. These ASICs and neuron-like digital chips appear to be the wave of the near future.

2.6.3 BENEFITS OF ARTIFICIAL NEURAL NETWORKS (ANN)

Neural networks gets computing power through its parallel distributed structure and its ability to learn and generalize, producing reasonable output for input not encountered during learning. Neural networks cannot provide solution by working individually, but they need to be integrated into a consistent system. Neural networks offer the following properties and capabilities [103]:-

1. ***Nonlinerity*** – An artificial neuron can be linear or nonlinear. A neural network made up of interconnected nonlinear neuron is itself nonlinear. The nonlinearity is distributed throughout the network. Nonlinearity is important if the input signal is nonlinear [107,108].

2. ***Input-Output Mapping*** - Supervised learning involves the modification of the synaptic weights of a neural network by applying a set of labeled training samples. Example consists of unique input signal and corresponding desired response. The neural network is presented an example and the synaptic weights of the neural networks are modified to minimize the difference between the desired response and the actual response produced by the input signal. Training of the neural network is repeated until the neural network reaches a steady state where there are no further significant changes in the synaptic weights. Thus, the neural network learns from examples by constructing an input-output map.

3. ***Adaptively*** - Neural networks have the capability to adapt their synaptic weights to changes in the surrounding environment [105]. Neural networks can be easily retrained to deal with minor changes [108]. In a nonstationary environment, the neural network can change its synaptic weights in real time. A system that is adaptive is not always robust. An adaptive system with short time constants may change rapidly and tend to respond to spurious disturbances, causing degradation in system performance. The time constant of the system should be long enough for the system to ignore spurious disturbance and yet be short enough to respond to meaningful changes for the system to be fully adaptive in the environment.

4. ***Evidential Response*** – In pattern classification, a neural network can be designed not only to provide information about which particular pattern to select, but also about the confidence in the decision made. The latter information may be used to reject ambiguous patterns, thus improving performance of the network.

5. ***Contextual Information*** – Knowledge is represented by the structure and activation of the neural networks. Every neuron is potentially affected by the global activity of all the other neurons in the neural networks. Thus, contextual information is natural in neural networks.

6. ***Fault Tolerance*** - A neural network in the hardware form, has the potential to be inherently fault tolerant. Its performance degrades gracefully under adverse operating conditions [105]. If a neuron or its connecting links are damaged, recall

of a stored pattern is impaired in quality. The damage has to be extensive before overall response of the neural networks degrades seriously.

7. ***VLSI Implementability*** – The parallel nature of neural networks makes it faster for certain computation task [108]. This also makes a neural network suited for implantation using VLSI technology.

8. ***Uniform of Analysis and Design*** – Commonly, neural network is an information processor. The same notation is used in all domains involving the application of neural networks [107].

9. ***Neurobiological Analogy*** – The brain motivated the design of neural networks. Neurobiologists use neural networks as a research tool for the interpretation of neurobiological phenomena. Engineers use new neurobiology ideas to solve more complex problems.

2.6.4 DIFFERENCE BETWEEN ARTIFICIAL NEURAL NETWORKS (ANN) AND TRADITIONAL COMPUTING

Neural networks offer a different way to analyze data and to recognize patterns within data. However, they are not a solution to all computing problems. Traditional computing methods work well for problems that can be characterized. Table 2.5 identifies the basic similarities and differences between neural networks and traditional computing [104,109].

Table 2.5 – Difference Between Traditional Computing and ANN

Characteristic	*Traditional Computing*	*ANN*
Processing	Sequential, digital and synchronous	Parallel
Functions	Logically using rules, concepts and calculations	Gestural using images, pictures, controls
Learning Method	Programmed with instructions	Trained using examples
Memory	Memory and processing separate	Memory and processing elements are collacated
Fault-Tolerant	Not fault-tolerant	Maybe fault-tolerant
Knowledge	Knowledge stored in an addressed memory location is strictly replaceable	Knowledge is adaptable; information is stored in interconnection between neurons
Processing	Autocratic	Anarchic
Cycle time	In nanoseconds	In miliseconds
Software	Software-dependent	Self-organizing during learning
Application	Accounting, word processing, math, inventory, digital communications	Sensor processing, speech recognition, pattern recognition, text recognition

Traditional computers are ideal for many applications. They can process data, track inventories, network results, and protect equipment.

ANN offer a different approach to problem solving and are sometimes called the sixth generation of computing. It tries to provide a tool that programs itself and learns on its own. Neural networks provide the capability to solve problems without the benefits of an expert and without the need of programming [104]. It can seek patterns in data that no one knows are there.

Despite the advantages of neural networks over traditional computing in these specific areas, it is not a complete solution. It learns, and as such, it does continue to make "mistakes." Furthermore, there is no way to ensure that the network is the optimal network.

Neural networks requires a designer to meet a number of conditions, which include [104]:

- A data set which includes the information which can characterize the problem.
- An adequately sized data set to both train and test the network.
- An understanding of the basic nature of the problem
- An understanding of the development tools.
- Adequate processing power

Neural networks offer the opportunity of solving problems in an arena where traditional processors lack both the processing power and a step-by-step methodology

[104]. A number of very complicated problems cannot be solved in the traditional computing environments. Examples are in the field of speech recognition and image recognition.

2.6.5 ANALOGY OF THE BRAIN WITH ARTIFICIAL NEURAL NETWORK (ANN)

The basic element of the human brain is a specific type of cell known called neuron. Since, neurons do not regenerate, it is assumed that these cells provide us with the human ability to remember, think, and apply previous experiences [104]. The human brain has 100 billion neurons and each neuron is typically connected to 1,000 to 10, 000 neurons [104,103]. Figure 2.23 shows a neuron.

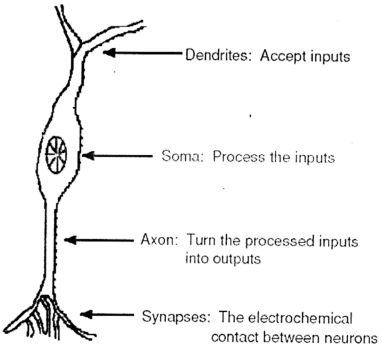


Figure 2.23 : Model of a Neuron

Synapses are elementary structural and functional units that mediate the interaction between neurons. A synapse is a simple connection that can impose or inhibition excitation, but not both on the receptive neuron [103]. In an adult brain, plasticity is accounted for 2 mechanisms : the creation of new synaptic connection and the modification of existing synapses. Axons are transmission lines and dendrites are the receptive zones [103,104]. Soma processes the incoming signals over time into an output which is sent to other neurons through axons and synapses [104].

Biological neurons are significantly more complex than the simple explanation above. They are also significantly more complex than the existing artificial neurons. However, network designers continue to improve their systems as biology provides a better understanding of neurons and as technology advances [104,103].

2.6.6 MODEL OF A NEURON

A neuron is an information processing unit that is the fundamental to the operation of a neural network. Figure 2.24 shows the model of a neuron that forms a basis for designing neural networks.

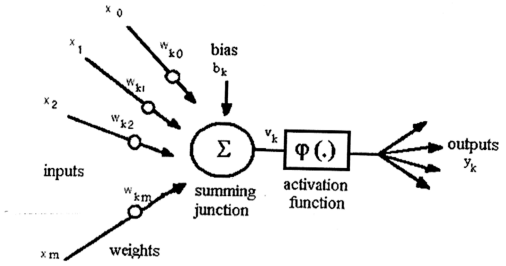


Figure 2.24 - Nonliear Model of a Neuron

The basic elements of a neuron are :-

1. Synapses is characterized by a weight or strength of its own. A signal x_j at the input of the synapses *j* connected to neuron *k* is multiplied by the synaptic weight w_{kj}. The first subscript of w_{kj}, *k*, refers to the neuron in question and the second subscript, j, refers to the input end of the synapses [103,106]. The synaptic weight of an artificial neuron lie in a range that include negative as well as positive values.

2. An adder for summing the input signals, weighted by the respective synapses of the neuron [103,106].

3. An activation function for limiting the amplitude of the output of a neuron. It is also referred as a squashing function [107,103,106]. The normal amplitude range of the output of an neuron is [0,1] or [-1,1].

4. An external bias, b_k, has the effect of increasing or lowering the net input of the activation functions [103,106], depending on weather it is positive of negative, respectively.

A neuron k, is described by Equation 2.1 and Equation 2.2.

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \qquad \text{Equation 2.1}$$

$$yk = \varphi(u_k + b_k) \qquad \text{Equation 2.2}$$

where

$x_1, x_2, \ldots, x_m$ are the input signals

$w_{k1}, w_{k2}, \ldots w_{km}$ are the synaptic weights of neuron k

u_k is the linear combiner output

b_k is the bias

$\varphi(.)$ is the activation function

y_k is the output signal of neuron k

The use of bias, b_k has the effect of applying an affine transformation to the output u_k shown by Equation 2.3.

$$v_k = u_k + b_k \qquad \text{Equation 2.3}$$

Activation function, $\varphi(v)$, defines the output of a neuron in terms of local field v. Three basic types of activation functions are [110,103,111,106,100]:-

1. ***Threshold Function*** – This type of activation function is shown in Figure 2.25. It is defined in Equation 2.4. The output neuron takes the value 1 if the induced field of that neuron is non-negative and 0 otherwise.

$$\varphi(v) = \begin{cases} 1 \text{ if } v > 0 \\ 0 \text{ if } v \leq 0 \end{cases} \qquad \text{Equation 2.4}$$

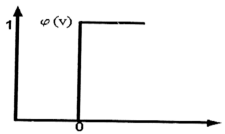


Figure 2.25 – Threshold Transfer Function

2. ***Piecewise Linear Function*** – This type of function is shown in Figure 2.26.. It is defined in Equation 2.5. The amplification factor inside the linear region of operation is assumed to be unity.

$$\varphi(v) = \begin{cases} 1 & \text{if} \quad v \geq +1/2 \\ v & \text{if} \quad +1/2 > v > -1/2 \\ 0 & \text{if} \quad v \leq -1/2 \end{cases} \qquad \text{Equation 2.5}$$

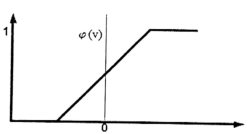


Figure 2.26 – Piecewise Linear Function

3. ***Sigmoid Function*** – The sigmoid function graph is s-shaped (Figure 2.27). It is the most common form of activation function used in neural networks. An example of the sigmoid function is the logistic function defined in Equation 2.6 where a is the slope parameter. Activation functions such as *tanh* that produce both positive and negative values tend to yield faster training than activation functions such as logistics that produce only positive values, because of better numerical conditioning [112].

$$\varphi(v) = \frac{1}{1 + \exp(-av)} \qquad \text{Equation 2.6}$$

Figure 2.27 – Sigmoid Function

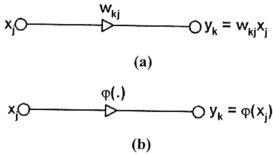


Figure 2.28 – Signal Flow Graphs

A signal flow graph is a simplified appearance of a neuron [103]. It is a network of directed links that are interconnected at certain points called nodes. A node j has an associated node signal x_j. A directed link originates at node j and terminates on node k. It has an associated transfer function or transmittance that specified the manner in which the signal y_k at the node k depends on the signal x_j at node j. The signal flow of a graph is dictated by these rules [103]:

1. A signal flows along a link in the direction directed by the arrow on the link.

   There are two different types of links

   - Synaptic links which are governed by a linear input-output relation. The node signal x_j is multiplied by the synaptic weight w_{kj} to produce the signal y_k, as shown in Figure 2.28(a).
   - Activation links which are governed by a non-linear input-output relation. This relationship is illustrated in Figure 2.28(b), where $\varphi(.)$ is the non-linear activation function.

2. A node signal equals the algebraic sum of all the signals entering the pertinent node via the incoming links. This is illustrated in Figure 2.28(c).
3. The signal at a node is transmitted to each outgoing link originating from that node, with the transmission being entirely independent of the transfer function of the outgoing link. This is illustrated in Figure 2.28(d).

Figure 2.29 gives a signal flow graph of a neuron, k. It is simpler the the neuron model in Figure 2.24, and it contains all the functional details.

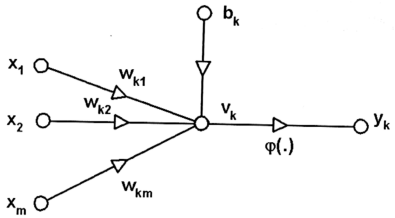


Figure 2.29 – Signal Flow Graph of a Neuron

2.6.7 ARTIFICIAL NEURAL NETWORK (ANN) ARCHITECTURES

There are three fundamental different classes of network architectures [103].

1. Single Layer Feedforward Network

In a single layer network, there is input layer of source nodes that projects onto the output layer of neurons but not vise versa. It is a feedforward or acyclic type network. Figure 2.30 illustrates a four-node single layer feedforward network.

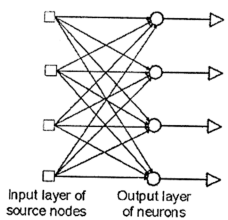


Figure 2.30 – Single Layer Feedforward Network

2. Multilayer Feedfoward Network

A multilayer feedforward network has one or more hidden layers, which consist of hidden neurons. The hidden neurons intervene between the input and the output in some useful manner. This enables the network to extract higher order statistics. The architectural graph in Figure 2.31 illustrates the layout of a multilayer feedforward neural network. The network in Figure 2.31 is referred as a 6-4-2 network because it has 6 source nodes, 4 hidden neurons and 2 output neurons.

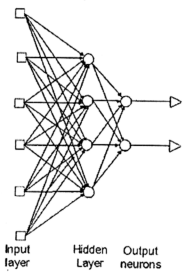


Figure 2.31 – Feedforward NN with One Hidden Layer

3. **Recurrent Network**

A recurrent network has at least one feedback loop. In Figure 2.32, a recurrent network consists of a single layer of neurons with each neuron feeding its output signal back to the inputs of all the other neurons. This network has no self feedback loops or hidden layers. The feedback loops use a unit-delay element, which results in a non-linear dynamic behavior.

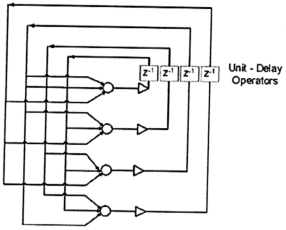


Figure 2.32 – Recurrent NN with No Self-Feedback and No Hidden Neurons

2.6.8 TRAINING AN ARTIFICIAL NEURAL NETWORK (ANN)

Neural networks have two training modes – learning and retrieval. A network is trained after it is structured for an application. Learning or training is the capacity to absorb information from its environment without requiring some external intelligent agent to "program" it [113,103]. During training, data is loaded permanently into the memory base [103]. Retrieval is the process when associative data is recalled from memory [103]. Figure 2.33 depicts this paradigm.

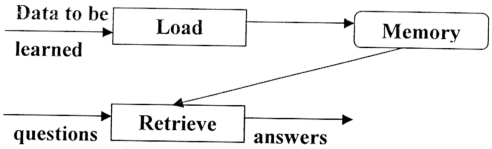


Figure 2.33 – A Simple Model of Learning

The initial weights are chosen randomly, when the learning or training begins. There are two approaches to training – supervised and unsupervised. Supervised training provides the network with a desired output, through "grading" the network performance or by provides the desired outputs with the inputs. Unsupervised training tries to make sense of the inputs without outside help.

Supervised Learning – In supervised training, the input and output are provided [103,104]. The neural networks will process the input and compare its output with the desired output [103,104]. Errors are propagated back, causing the system to adjust its weights. This process is repeated over and over again until the connection weights are refined. The data set used for training is called the "training set". Neural networks are monitored to determine if the system is simply memorizing its data in some nonsignificant way [104]. A set of data is held back to test the system after training.

Training phase is time consuming [104]. Training is considered complete when the neural networks reached the user defined performance level. This level signifies that the network has achieved the desired statistical accuracy as it produces the required output for a given sequence of input. The weights are frozen when no further learning is required. Certain neural networks continue training at a slower rate while in operation. This would help the neural networks to adapt to the gradually changing conditions.

Training sets have to be large to contain all the needed information for the neural network to learn the features and relationships that are important [103]. The training sessions also must include a variety of data. If the neural network is trained with only one set of data, all the weights that was set would be drastically altered when learning the next fact [103]. The previous fact could be forgotten causing the system to learn everything together, and finding the best weight setting for the total set of facts.

The representation of the input and output data is also important to successfully train a neural network. Raw data from external environment must be converted to the desired input of the neural networks. Data are also scaled or normalized to the network paradigm.

When a network is unable to solve a problem, the designer has to review these parameter [104]:-

- The inputs
- The outputs
- The number of layers
- The number of elements per layer
- The connections between the layers
- The summation function
- The transfer function
- The training function
- The initial weights

Unsupervised Learning - In unsupervised training the network is provided with the input but without the desired output [103]. The system will decide on its own what feature to use to group the input data. There are two types of unsupervised training, reinforcement learning and self-organized learning [103].

In reinforcement learning, the learning of an input-output mapping is performed through continued interaction with the environment in order to minimize a scalar index of performance [103]. Reinforcement learning is related to dynamic programming. Dynamic programming provides the mathematical formalism for sequential decision making.

In self-organized learning, there is no external teacher to over see the learning process. An independent measure of the quality of representation that the neural network is required to learn and the free parameters of the neural networks are optimized with respect to that measure [103]. Self-organized learning is performed using a competitive learning rule. For example : A neural networks that consist of two layers, an input layer and a competitive layer. The input layer receives the available data, while the competitive layer consists of neurons that compete with each other for the opportunity to respond to features contained in the input data [103]. The neural networks could operate in accordance with the "winner-takes-all" strategy.

Several factors are taken into consideration when training a neural network [103,104]:

- Time
- Network complexity
- Size
- Paradigm selection
- Architecture
- Type of learning rule
- The desired accuracy

The factors above have a significant role in determining how long it will take to train a network. Changes in any of these factors may extend the training time to an unreasonable length or result in unacceptable accuracy. The learning rate of a neural network is usually between zero and one. If the learning rate is greater than one, the learning algorithm might have overshot in correcting the weights and the neural networks will oscillate [104,103]. Small learning rate will not correct the current error quickly. If small steps are taken in correcting errors, the best minimum convergence can be achieved.

There are many learning rules used in neural networks. Most rules have some variation of the oldest learning law, Hebb's Rule. However, researchers always bring up new ideas. But the understanding of how neural processing works is limited. Learning is more complex that the simplifications represented by the learning law. Some of the major learning laws are :-

Hebb's rule was introduced by Donald Hebb in 1949. In this rule, if a neuron receives an input from another neuron, and if they are highly active, then the weight between the neuron would be strengthened [102].

Hopfield law is similar to Hebb's rule. This rule also states the specific magnitude of the strengthening or weakening. It states that if the desired output and the input are either active or inactive, increment the connection weights by the learning rate, otherwise, decrement the weights by the learning rate [104].

The delta rule is a variation of the Hebb' rule. It is also known as the error correction learning rule [101]. It is one of the most commonly used rules. In this rule, the strengths of the input connections is continuously modified to reduce the difference between the desired output value and the actual output of the processing element. It changes the synaptic weights to minimize the mean square error of the neural network [102]. It is also called the Widrow-Hoff learning rule and the least mean square learning rule [104].

The delta error in the output layer is transformed by the derivative of the transfer function and is then used in the previous layer to adjust the input connection weights. The error is back-propagated into the previous layers. This process continues until the first layer is reached [104]. When this rule is used, it is important to randomize the input data. Well ordered training sets leads to a network that does not converge to the desired accuracy, which makes the network incapable of learning the problem.

The gradient descent rule is similar to the delta rule in the derivative of the transfer function used to modify the delta error before applying it to the connection weights [104]. However, an additional proportional constant is tied to the learning rate is appended to the final modifying factor acting on the weight. It converges to stability very slowly. Different learning rates for different layers help the neural network to stabilize faster [104]. Learning rates of layers closer to the output are usually set lower than the layers nearer the input.

Kohonen's learning law was developed by Teuvo Kohonen. The learning elements compete for the opportunity to learn or update their weights. The processing element with the largest output is declared the winner and has the capability of inhibiting or exciting its competitors [104,103]. Only the winner is permitted an output, and only the winner plus its neighbors are allowed to adjust their weights. The size of the neighborhood also can vary during training. This usually starts with a large definition of neighborhood, and narrowed during training. The winning element is defined as the one with the closest match with the input pattern 103]. This rule is good for statistical or topological modeling of data. It is sometimes referred as self organizing map or self-organizing topology.

Memory based learning rule is also known as the nearest neighbor rule [103]. In this method, all of the past experiences are explicitly stored in a large memory of correctly classified input-output examples [103]. All memory based learning algorithms have two ingredients [103]:-

- Criterion used for defining the local neighborhood of the test vector.

- Learning rule applied to the training examples in the local neighborhood of the test vector.

The local neighborhood is defined as the training example that lies in the immediate neighborhood of the test vector.

Boltzman learing rule was named in the honor of Ludwig Boltzmann. It is a stochastic learning algorithm derived from statistical mechanism [103]. In a Boltzman machine the neurons constitute of a recurrent structure, which operate in a binary manner. The neurons are either "on" or "off". The machine operates by choosing a neuron at random at some step of learning process. It flips the state of the neuron at some pseudotemperature. When the rule is applied repeatedly, the machine will reach thermal equilibrium [103].

2.6.9 USES OF ARTIFICIAL NEURAL NETWORKS (ANN)

1. ***Language Processing*** - Language processing encompasses a variety of applications, such as text-to-speech conversion, auditory input for machines, automatic language translation, secure voice keyed lock, automatic transcription, aid for the deaf, aid for the physically disabled which respond to voice commands, and natural language processing [105]. Many companies and universities are researching how computers, via ANNs, could be programmed to respond to spoken commands. The speech-parsing system by Apple Corporation, can recognize almost any person's speech through a limited vocabulary. Neural

networks are also used for speech generation and speech recognition [109].

2. ***Character Recognition*** - Character recognition is another area in which neural networks are providing solutions [109]. HNC Inc. markets a neural network based product that can recognize hand printed characters through a scanner. It is 98% to 99% accurate for numbers, a little less for alphabetical characters. QNspec has the capability of recognizing characters, including cursive handwriting [105].

3. ***Data Compression*** - A number of studies have been done proving that neural networks can perform real-time compression and decompression of data. These networks are auto associative in that they can reduce eight bits of data to three and then reverse that process restructuring to eight bits again [109].

4. ***Pattern Recognition*** - Neural networks are used as a recognizer of patterns in the quality control field. A number of automated quality applications are currently in use. These applications are designed to find that one in a hundred or one in a thousand part that is defective. Neural networks are also used in sensor processors. Many of these sensor-processing applications exist within the defense industry. The neural networks systems have shown success at recognizing targets. The sensor processors take data from cameras, sonar systems, seismic recorders, and infrared sensors and then use the data to identify probable phenomenon.

5. ***Signal Processing*** - Neural networks have proven capable of filtering noise. Widrow's MADALINE eliminates noise from phone lines. Neural networks are

also used to regenerate analogous signals after transmission on a defective channel [105].

6. ***Time Series Prediction*** - Banking, credit card companies, and lending institutions deal with decisions that are not clear cut which involve learning and statistical trends [107]. The loan approval process involves filling out forms which can enable a loan officer to make a decision. The data from these forms is now being used by neural networks which have been trained on the data from past decisions [107]. Neural networks are also used in the financial markets – stock, bonds, international currency, and commodities [105]. Neural networks have been reported to be highly successful in the Japanese financial markets. Attasoft PredictorPro commercial software is used to predict the Stock Market [114].

7. ***Servo Control*** - Controlling complicated systems is one of the more promising areas of neural networks [107]. They offer two advantages. Firstly, the statistical model of neural networks is more complex enabling it to handle a wider variety of operating conditions without having to be retuned. Secondly, because neural networks learn on their own, they do not require control system's expert. In the oil industry, a neural network has been applied to the refinery process. The network controls the flow of materials and is touted to do it more vigilantly than humans.

8. ***Robotics*** – Neural networks have innovated the potential that allow robots to learn and adapt to the environment. Neural networks provide a new mode of implementation and a more effective solution for industrial applications. Neural

networks are used to deal with the navigation of mobile robots [106]. Neural networks learn from a constant set of obstacles and the system is trained in escape strategies [105].

9. ***Optimization*** – Neural networks provide solutions for optimization problems such as the "Traveling Salesman Problem" [109]. The Bolzman machine is used in the optimization of waiting time of flight crew between connecting flights [105].

10. ***Image Processing and Computer Vision*** - In the medical field, neural networks are used to identify and segment meaningful objects from medical images [111]. Attasoft ImageFinder software does content based image segmentation [114]. ANN are also used in image matching, preprocessing and analysis, computer vision, stereo vision, and processing and understanding of time-varying images [100].

2.6.10 SELF ORGANIZING MAP (SOM) NEURAL NETWORKS

Self organizing map (SOM) finds natural clusters or features similarities from unlabeled training data [108]. SOM was developed by Teuvo Kohonen of Helsinki University in 1982 [108,115]. SOM is based on competitive learning [103]. The output neurons of the neural network competes among themselves to be activated or fired, with the result that only one output neuron is on at any given time [109,108]. The output neuron that wins the competition is called a winner-takes-all neuron or a

winning neuron. The connection to the winning unit and connections to units in its neighborhood are modified [107,109,110]. Winner-takes all competition is induced among the output neurons by using lateral inhibitory connections between them [106,103,108]. Example of SOM applications include data compression [106], feature extraction, preprocessing weights or data for other neural networks [108], data analysis, pattern recognition, speech analysis, robotics, industrial and medical diagnostics, instrumentation and control [115]

In a SOM, the neurons are placed in a 1D or 2D lattice [103]. A self organizing map is characterized by the formation of a topographic map of the input pattern in which the spatial location of the neurons in the lattice are indicative of intrinsic statistical features contained in the input patterns [115,108].

The development of SOM is motivated by a distinct feature of the human brain [103,112]. The brain is organized in many places in such a way that different sensory input are represented by topologically ordered computational maps [115] [116]. The sensory inputs such as tactile, visual and acoustic are mapped onto different areas of the cerebral cortex in a topological ordered manner [116]. The computational map constitutes a basic building block in the information processing infrastructure of the nervous system. The neurons transform input signals into a place-coded probability distribution that represent the computed values of parameters by site of maximum relative activity within the map [108,117].

The goal of the SOM is to transfer an incoming signal pattern of arbitrary dimension

into a 1D or 2D discrete map and to perform this transformation adaptively in a topologically ordered fashion [117,116,103,108]. Figure 2.34 shows the schematic diagram of a 2D lattice of neurons used as discrete map. Each neuron in the lattice is fully connected to all the source nodes in the input layer. This neural network represents a feedforward structure with a single computational layer consisting of neurons arranged in rows and columns [103]. A 1D lattice consists of a single column or row of neurons.

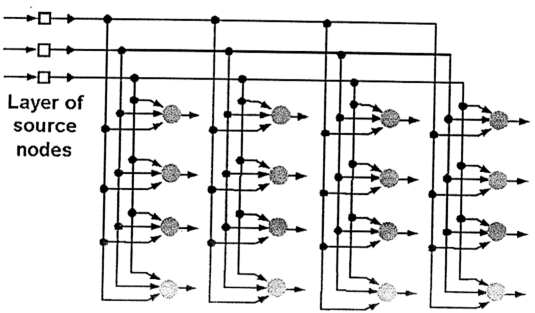


Figure 2.34 – 2D Lattice of Neurons [103,100]

Each input pattern to the neural networks consists of a localized region of activity against a quiet background. The location and nature of a spot varies from one input pattern to another. All the neurons in the neural networks are exposed sufficiently to a number of different realizations of the input pattern to ensure that the neural network can mature properly.

The weights of the SOM are initialized by assigning them with small random values. After initialization process, three processes are involved in the formation of the SOM neural network [103]:-

1. ***Competition*** – The neurons in the network compute their respective values of discriminate function for each input pattern. The discriminate function provides the basis for competition among the neurons. The neuron with the largest value of discriminate function is declared as the winner of the competition.

2. ***Cooperation*** – The winning neuron determines the spatial location of a topological neighborhood of excited neurons, providing the basis for cooperation among neighboring neurons [118,103,109]. The winning neuron can update its parametric weight vectors using two strategies [119]
   - The learning rate strategy. All other neurons keep their old values
   - Rewards and punishment strategy. Neurons near the winning neuron are positively updated while neurons farther away are updated negatively.

Figure 2.35 demonstrates the concept of neighborhood with the radius one surrounding neuron 13. The neighborhood neurons for neuron 13 are neuron 8, 12, 13, 14 and 18 [118].

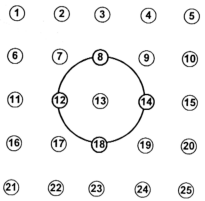


Figure 2.35 – Neighborhoods

3. ***Synaptic Adaptation*** - This mechanism enables the excited neurons to increase their individual values of the discriminate function in relation to the input pattern through suitable adjustments applied to their synaptic weights. The adjustments are made such that the response of the winning neuron to the subsequent application of a similar input pattern is enhanced.

2.7 FUZZY LOGIC

Figure 2.36 illustrates the flow of Section 2.7. Section 2.7.1 gives an introduction to fuzzy logic. Fuzzy set is described in Section 2.7.2. Various methods to obtain membership functions are explained in Section 2.7.3. Fuzzy logic operators are given in Section 2.7.4. Several fuzzy to crisp conversion methods are illustrated in Section 2.7.5. The benefits and applications of fuzzy logic are described in Section 2.7.6 and Section 2.7.7 respectively. Fuzzy clustering is explained in detail in Section 2.7.8.

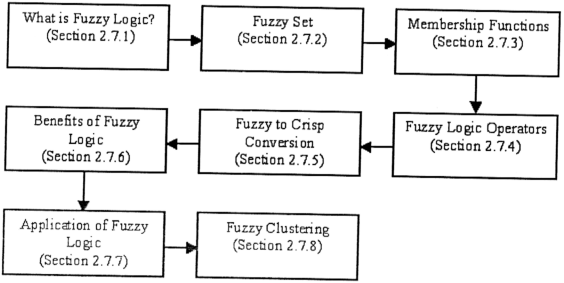


Figure 2.36 – Overview of Section 2.7

2.7.1 WHAT IS FUZZY LOGIC?

After the introduction of fuzzy logic by Lotfi Zadeh in 1965 [120,121,122], there were many theoretical developments in this field. However, Japanese researchers were a primary force in the commercializing of the fuzzy logic theory [120,121], with over 2000 patents.

Fuzzy logic provides a mean to represent uncertainties. It is a tool for modeling uncertainties associated with vagueness, imprecision, and/or lack of information regarding a particular element of the problem at hand [120]. Fuzzy logic does a good job trading off between significance and precision, which humans do with ease [122]. Figure 2.37 pictures the difference between precision and significance.

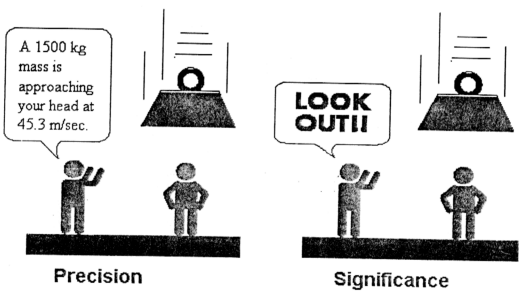


Figure 2.37 -- Precision and Significance

2.7.2 FUZZY SET

Fuzzy set is a set without crisp, clear, defined boundary [122]. It can contain elements with only a partial degree of membership. Classical set is a container that exclusively includes or wholly excludes any given element [122]. For example, the days of a week are Sunday to Saturday. These are wholly inclusive in the set of "days of the week". Other things like shoe polish, dorsal fins, liberty and butter are wholly excluded from the set. Figure 2.38 illustrates this example. "Days of the week" is called classical set. Classical sets follow the Laws of the Excluded Middle, which states that X must either be in set A or in a set not-A. Aristotle formulated this law.

Figure 2.38 – Days of the Week

This set is now compared with the set called "days of the weekend". It is agreed that Monday to Thursday are working days while Saturday and Sunday are weekends. However, Friday sits on the fence. Individual perception and cultural background are taken into account when weekend is defined. Weekend can be defined as "the period of Friday night or Saturday to Monday morning". Figure 2.39 illustrates the set days of a weekend.

Figure 2.39 – Days of the Weekend

Fuzzy logic is useful because truth of any statement is becomes a matter of degree in fuzzy logic [122]. It has the ability to reply a yes-no question with a not-quite-yes-or-no answer. In Boolean logic, we can give the numerical value "1" for yes and "0" for no. It does not allow any in-between values. However fuzzy logic permits in-between values like 0.2 or 0.836. Table 2.6 compares the Boolean value and fuzzy value of "weekend-ness" for each day in a week. For example, Friday has the fuzzy of value 0.8 which means that for most parts it is a weekend, while not completely. Figure 2.40 (a) and (b) are plots that shows the truth value for "weekend-ness" for Boolean logic and fuzzy logic respectively.

Table 2.6 – Boolean logic and fuzzy logic "weekend-ness" truth values

Days in a week	*Boolean Value*	*Fuzzy value*
Sunday	1	0.95
Monday	0	0
Tuesday	0	0
Wednesday	0	0
Thursday	0	0.3
Friday	0	0.8
Saturday	1	1

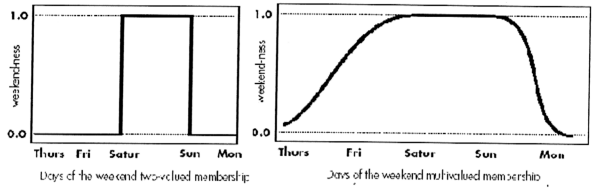


Figure 2.40 – "Weekend-ness" Truth Value (a) Boolean logic (b) Fuzzy Logic

In Figure 2.40(a), the weekend starts when the clock strikes 12 midnight on Friday, where the "weekend-ness" value jumps discontinuously from 0 to 1. Even through this is one way to represent the weekend, it does not connect with the real world. Figure 2.40(b) shows a smoothly varying curve of "weekend-ness". Friday and Thursday obtain partial membership in the fuzzy set of weekend. The curve that represents "weekend-ness" in Figure 2.40(b) is called a membership function.

Fuzziness describes the ambiguity of an event whereas randomness describes the uncertainty in the occurrence of the event [120]. An example of the difference is given as follows: Suppose a source states that a football player in Country A has a 95 percent chance of being over 7 feet tall. Another source states that a football player in Country B has a 0.95 membership in set of "very tall" people. The first source is a random quantity. There is a 5 percent chance that the player in Country A is not over 7 feet and could be someone who is extremely short. The second source contains less uncertainty because if the player turned out to be less than 7 feet tall, there is still a

high likelihood that he would be still be quite tall. The event will occur or not occur; but the description of the event is unambiguous enough to measure its occurrence or nonoccurrence.

2.7.3 MEMBERSHIP FUNCTIONS

A membership function (MF) is a curve that defines how each point in the input space (or universe of discourse) is mapped to a membership value (or degree of membership) between 0 or 1 [122]. An example of MF is shown in Figure 2.40(b) which gives the membership of each day of the week for the weekend set.

Fuzzy set of tall people is another common example. In classical a set, people taller 6 feet are defined as "tall" and "short" if they are shorter than 6 feet. Figure 2.41 gives the classical set plot.

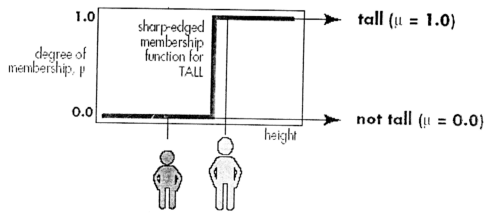


Figure 2.41 – Classical Set Plot for Tall People

However in the real world, it is unreasonable to call someone 6 feet as tall and someone else who is 5'11" as short. Fuzzy logic would give a membership function,

μ, as in Figure 2.42. It shows a smoothly varying curve that passes from not-tall to tall. The output-axis is a number known as the membership value between 0 and 1.

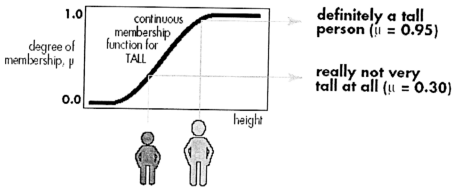


Figure 2.42 – Membership Function for Tall People

Subjective interpretation and appropriate units are built into the fuzzy set. The membership takes into account whether, "She is tall" refers to a small girl or a grown woman. The units are also included in the curve as it makes no sense to say "Is she tall in inches or in meters?".

Membership functions must be between the value of 0 and 1. The shape of a membership function can be a arbitrary curve that suits the application. A fuzzy set A in X is defined as a set of ordered pairs (Equation 2.7) [122]:-

$$A = \{x, \mu_A(x) \mid x \in X\} \qquad \text{Equation 2.7}$$

X is the universe of discourse

x are elements in X

$\mu_A(x)$ the MF of x in A

$\mu_A(x)$ maps each element of X to a membership value between 0 and 1. Figure 2.43 gives some common shapes of membership functions.

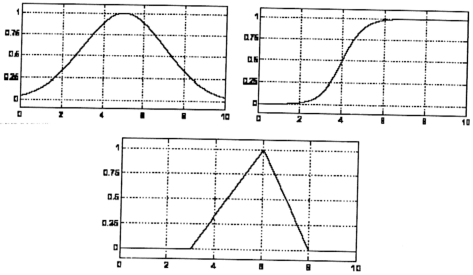


Figure 2.43 – Common Shapes of Membership Functions

There are many ways to assign membership function values or functions to fuzzy variables [120]:-

1. ***Intuitions*** – It derives from the capacity of humans to develop membership function through their own innate intelligence and understanding. Intuition involves contextual and semantic knowledge about an issue. The example shown in Figure 2.44 is the membership functions for the fuzzy variable temperature. Each curve is a membership function corresponding to various fuzzy variables such as cold, warm and hot.

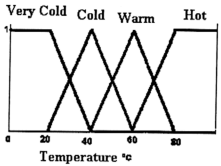


Figure 2.44 – Membership Function for Fuzzy Variable "Temperature"

2. ***Inference*** – In inference, knowledge and facts are used to perform deductive reasoning.

3. ***Rank Ordering*** – In rank ordering, preferences is assessed by a single individual, a committee, a poll or other opinion methods that can be used to assign membership values to a fuzzy variable. Preference is determined by pairwise comparisons, and these determine the ordering of the membership.

4. ***Angular Fuzzy Sets*** – Angular fuzzy sets are defined in a universe of angles, hence repeating shapes every 2 π cycle. It only differs from the standard fuzzy set in its coordinate description.

5. ***Artificial Neural Network (ANN)*** – ANNs also can be used to determine membership functions. ANNs are discussed in more detail in Section 2.6.

6. ***Genetic Algorithm*** – Genetic algorithm uses the concept of Darwin's theory of evolution. This theory is translated into algorithms to search for solutions in a more "natural" way. First, the different possible solutions are created, which are tested for their performance. A fraction of good solutions are then selected, and they undergo the process of reproduction, crossover, and mutation to create a new generation of solutions. This process is repeated until there is convergence within a generation. It searches for a solution from a broad spectrum of possible solutions, than restrict the search to a narrow domain. Genetic algorithms try to perform an intelligent search for a solution from an infinite number of possible solutions.

7. ***Inductive Reasoning*** – Membership functions can be generated using characteristics of inductive reasoning. The entropy is performed by entropy minimization principle, which clusters most optimally the parameters corresponding to the output class. This method generates membership functions based solely on the data provided. This method is quite useful for complex systems where the data is abundant and static. In situations where the data is dynamic, this method is not useful as the membership functions will continuously change with time. Three laws of induction are summarized as follows [120]:-

   - Given a set of irreducible outcomes of an experiment, the induced probabilities are those probabilities consistent with all available information that maximize the entry of the set.
   - The induced probability of a set of independent observation is proportional to the probability density of the induced probability of a single observation.

- The induced rule is that rule consistent with all available information of which the entropy is minimum.

The third rule is appropriate for classification (membership function development) and the second rule is appropriate for calculating the mean probability for each step of partition.

The goal of entropy minimization analysis is to determine the quantity of information in a given data set. The entropy of a probability distribution is a measure of the uncertainty of the distribution. This information measure compares the contents of data to aprior probability for the same data. The higher the prior estimate of the probability for an outcome to occur, the lower will be the information gained by observation is to occur. The entropy on a set of possible outcomes of a trial where one and only one outcome is true is defined by the summation of probability and the logarithm of the probability for all outcomes. In other words, the entropy is the expected value of information.

In a 1D case, lets assume that the probability of the *i*th sample w_i to be true is $\{p(w_i)\}$. When the sample w_i is actually observed in the future, and it is discovered as true, then Equation 2.8, is gained :

$$I(w_i) = -k \ln p(w_i)$$

Equation 2.8

where k is the normalizing parameter.

If w_i is discovered as false, the Equation 2.9 is gained.

$$I(\bar{w}_i) = -k\ln[1 - p(w_i)] \qquad \text{Equation 2.9}$$

The entropy of the inner products of all the samples (N) is given in Equation 2.10.

$$S = -k\sum_{i-1}^{N}[p_i \ln p_i + (1 - p_i)\ln(1 - p_i)] \qquad \text{Equation 2.10}$$

where $p_i = p(w_i)$.

The third rule of induction says that the entropy of a rule should be minimized. Minimum entropy, S, is associated with all p_is being as close to 1s and 0s as possible, which implies that they have a very high probability of either happening or not happening, respectively. This method is discussed in detail in Section 3.4.1.

ANN and generic algorithm make use of associated rules in the database to determine the membership function. Inductive reasoning produces good result when the database is not dynamic. When the database changes, the partitioning is reaccomplished. Inductive reasoning does not require convergence analysis, which makes it computationally less expensive compared to ANN and generic algorithm. Inductive reasoning uses the entire database to formulate the rules, which makes it computationally expensive when the database is large. The method used depends solely on the problem type and size.

2.7.4 FUZZY LOGICAL OPERATORS

Fuzzy logic is a superset to standard Boolean logic [122]. Standard logic can be used for fuzzy values at their extremes of 1 and 0. The minimum (A, B) function can resolve the standard A AND B statement, where A and B are limited to the range [0,1]. Table 2.7 gives the truth table for the AND operation and the corresponding fuzzy minimum operation. Figure 2.45(a) shows the AND operation of Boolean logic, while Figure 2.45(b) shows a graph of two fuzzy sets applied together using the minimum function to create one fuzzy set. The fuzzy minimum function is also known as fuzzy intersection or conjunction [122].

A	*B*	*A AND B*	*minimum (A,B)*
0	0	0	0
0	1	0	0
1	0	0	0
1	1	1	1

Table 2.7 – Truth Table For AND Operation and Minimum Function

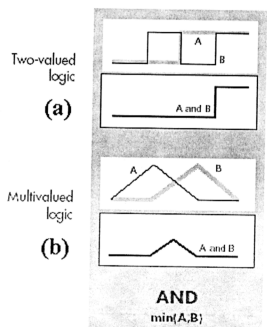


Figure 2.45 – AND Operation Graph Plot (a) Boolean AND Operation (b) Fuzzy Minimum Function

The OR operation is also replaced with the maximum function. Table 2.8 gives the truth table for OR operation and maximum function. Figure 2.46(a) shows a OR operation of Boolean logic, while Figure 2.46(b) shows how the maximum operation works over continuously varying range of truth values. The fuzzy maximum function is also known as fuzzy union or disjunction [122].

A	*B*	*A OR B*	*maximum (A,B)*
0	0	0	0
0	1	1	1
1	0	1	1
1	1	1	1

Table 2.8 – Truth Table for OR Operation and Maximum Function

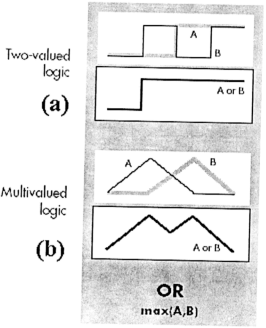


Figure 2.46 –OR Operation Graph Plot (a) Boolean OR Operation (b) Fuzzy Maximum Function

The NOT A operation becomes equivalent to the operation 1-A. Table 2.9 gives the truth table for NOT A operation and 1-A operation. Figure 2.47(a) shows a NOT operation of Boolean logic, while Figure 2.47(b) shows the NOT operation on fuzzy sets. The 1-A operation is also known as fuzzy complement [122].

A	*NOT A*	*1 - A*
0	1	1
1	0	0

Table 2.9 – Truth Table for NOT Operation and 1-A Function

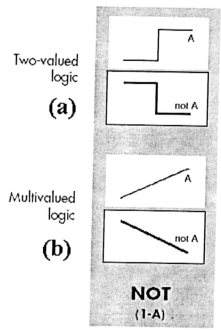


Figure 2.47 – NOT Operation Graph Plot (a) Boolean NOT Operation (b) Fuzzy NOT Function

2.7.5 FUZZY TO CRISP CONVERSIONS

Defuzzification of a fuzzy set is the process of "rounding it off" from its location in the unit hypercube to the nearest vertex. It reduces a fuzzy set to a crisp single-valued quantity. Defuzzification is the conversion of a fuzzy quantity to a precise quantity, just as fuzzification is the conversion of a precise quantity to fuzzy quantity [122]. The output of a fuzzy process can be the logical union of two or more fuzzy membership functions. For example, a fuzzy output comprises of two parts : a trapezoidal shape, C_1, (Figure 2.48(a)) and a triangle, C_2, (Figure 2.48(b)). The union of these two membership function, $C = C_1 + C_2$, involves the multi-operator, which is graphically the outer envelope of the two shapes, resulting in the shape shown in Figure 2.48(c).

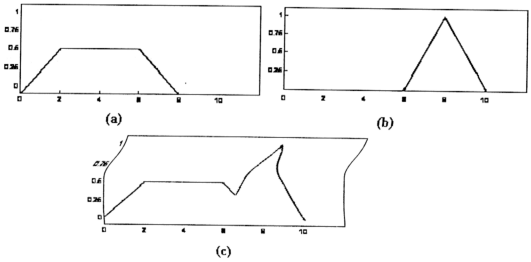


Figure 2.48 – Union of Two Membership Function

There are seven popular defuzzification methods mentioned in [120] :-

1. ***Max – Membership Principle Method*** - This method is also known as the height method. It is given by the algebraic expression in Equation 2.11.

$$\mu C(z^*) \geq \mu C(z) \quad \text{for all } z \in Z \qquad \text{Equation 2.11}$$

Max-membership principle is shown in Figure 2.49.

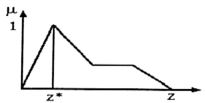


Figure 2.49 – Max-membership Defuzzification Method

2. ***Centroid Method*** – This method is also known as the center of area or center of gravity method. It is given by the algebraic expression in Equation 2.12. It favors central values in the universe of discourse, and could lead to slow inference cycle [110].

$$z^* = \frac{\int \mu_C(z).z dz}{\int \mu_C(z) dz} \qquad \text{Equation 2.12}$$

Centroid method is shown in Figure 2.50.

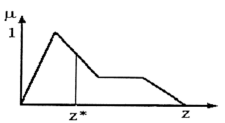


Figure 2.50 – Centroid Defuzzification Method

3. ***Weighted Average Method*** – This method is only valid for symmetrical output membership functions. It is formed by weighting each membership function in the output by its respective maximum membership value. It is given by the algebraic expression in Equation 2.13.

$$z^* = \frac{\sum \mu_C(Z).(Z)}{\sum \mu_C(Z)} \qquad \text{Equation 2.13}$$

From the example in Figure 2.51, the defuzzified value, z*, is calculated as in Equation 2.14.

$$z^* = \frac{a(0.5) + b(0.9)}{0.5 + 0.9} \qquad \text{Equation 2.14}$$

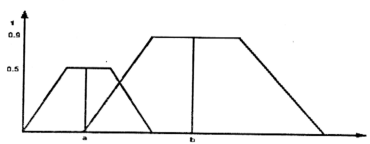


Figure 2.51 – Weighted Average Defuzzification Method

4. ***Mean – Max Membership Method*** – This method is also called the middle-of-maxima method. It is closely related to the max-membership principle, except that the location of the maximum membership can be non-unique. This method is given by the expression in Equation 2.15 where a and b are defined in Figure 2.52.

$$z^* = \frac{a+b}{2}$$ Equation 2.15

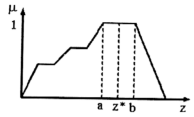


Figure 2.52 – Mean-Max Membership Defuzzification Method

5. ***Center of Sum Method*** – This method is faster than the other methods mentioned. It is similar to the centroid method, except in this method, the weights are the areas of the respective membership functions whereas in the centroid method the weights are individual membership values. The defuzzified area z* is given by Equation 2.16.

$$z^* = \frac{\int_z z \sum_{k=1}^{n} \mu_{Ck}(z) dz}{\int_z \sum_{k=1}^{n} \mu_{Ck}(z) dz}$$ Equation 2.16

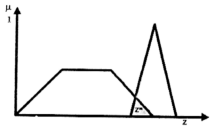


Figure 2.53 – Center of Sum Defuzzification Method

6. ***Center of Largest Area Method*** – If the output fuzzy set has at least two convex sub-regions, then the center of gravity of the convex fuzzy sub-region with the largest area is used to obtain the defuzzification value of the output. This is shown graphically in Figure 2.54 and given by the algebraical expression in Equation 2.17.

$$z^* = \frac{\int \mu_{C_m}(z) z dz}{\int \mu_{C_m}(z) dz} \qquad \text{Equation 2.17}$$

where C_m is the convex subregion that has the largest area making up C_k. This condition is applied in cases where the overall output C_k is nonconvex. In the case where C_k is convex, z^* is the same value as determined by the centroid method.

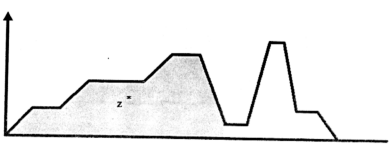


Figure 2.54– Center of Largest Area Defuzzification Method

7. ***First (or Last) of Maxima Method*** – This method uses the overall output or union of all individual output fuzzy sets to determine the smallest value of the domain with maximize membership degree. The equations for z^* are as follows. The largest height in the union (denoted $hgt(C_k)$ is determined in Equation 2.18.

$$hgt(C_k) = \sup_{z \in Z} \mu_{C_k}(z) \qquad \text{Equation 2.18}$$

Then the first of the maxima is found in Equation 2.19.

$$z^* = \inf_{z \in Z} \{ z \in Z | \mu_{C_k}(z) = hgt(C_k) \}$$

Equation 2.19

An alternative is the last of the maxima in Equation 2.20.

$$z^* = \sup_{z \in Z} \{ z \in Z | \mu_{C_k}(z) = hgt(C_k) \} \qquad \text{Equation 2.20}$$

Graphically this method is shown in Figure 2.55.

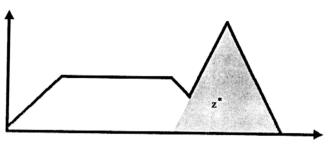


Figure 2.55 – First of Max (or Last of Max) Defuzzification Method

2.7.6 BENEFITS OF FUZZY LOGIC

1. Fuzzy logic has the ability to model highly complex and nonlinear problems [121].

2. Fuzzy rule based systems execute faster than conventional systems with fewer rules [121].

3. Fuzzy logic improves cognitive modeling of expert systems. Fuzzy systems have the ability to encode knowledge directly in a form very close to the way experts think [121]. Fuzzy logic also has the ability to model systems involving multiple experts.

4. Fuzzy logic reduces model complexity because it requires fewer rules than conventional systems [121]. The rules are expressed in natural language, closer to the way knowledge is expressed by humans.

5. Fuzzy logic improves the handling of uncertainties and possibilities [121].

2.7.7 APPLICATIONS OF FUZZY LOGIC

Fuzzy logic research has been conducted in the United States, Europe and Japan. However, Japanese researchers are the primary force in advancing the practical implementation of fuzzy logic, with over 2000 patents in this area [120].

1. ***Videography*** – Sony has made fuzzy logic camcorders that offer fuzzy focusing and image stabilization [120]. Omron has used fuzzy logic in the camera filming for the telecast of sporting events [123].

2. ***Transportation*** – Sendai, Japan has a 16 station subway that is controlled by a fuzzy computer. The ride is so smooth that riders do not need to hold straps and the controller makes 70% fewer judgment error that human controllers [120]. Nissan used fuzzy control for efficient and stable control of car-engines and for automobile cruise-control [123].

3. *House Hold Applicants* – Mitsushita built a fuzzy washing machine that combines smart sensor with fuzzy logic. The sensor detects the color and the kind of clothes present, and the fuzzy microprocessor selects the appropriate combination of water temperature, detergent amount, and wash and spin cycle time [120]. Mitsubishi and Sharp use fuzzy control systems to prevent unwanted temperature fluctuations in air-conditioning systems [123]. Japan has fuzzy golf diagnostic system, fuzzy toasters, fuzzy rice cookers, fuzzy vacuum cleaners and many more.

4. ***Finance*** – Tokyo's stock market has at least one stock-trading portfolio based on fuzzy logic that outperformed the Nikkei Exchange average.

5. ***Expert Systems*** – Fuzzy rules are used in expert systems to model the expert's knowledge [124]. Yamaichi and Hitachi have substitution of an expert for the assessment of stock exchange activities.

6. ***Clustering*** – Fuzzy clustering techniques are used for image and data clustering [125]. Kawasaki Medical School uses fuzzy clustering in medicine technology for cancer diagnosis [123].

7. ***Control Systems*** – Tokyo Electric Power had an automatic control of dam gates for hydroelectric powerplant using fuzzy control [123]. Hirota, Fuji Electric, Toshiba and Omron have developed simplified control of robots using fuzzy control system [123].

8. ***Documentation*** - Toshiba, Nippon-System and Keihan-Express use fuzzy logic for optimized planning of bus time-table [123]. Mitsubishi Electric use fuzzy logic in the archiving system of documents [123].

2.7.8 FUZZY CLUSTERING

Clustering algorithms are used to organize and categorize data, compress data and construct models [125]. Clustering partitions a data set into several groups that are similarity within a group 125]. Fuzzy clustering allows partitioning of objects in such a way that some object definitely belong to certain group, but other objects whose group membership is less evident [126,127]. Figure 2.56 depicts a set of objects described by two variables. Each point has a membership value which specifies the extent to which that object can be regarded as belonging to group G_1. High values denote a high level of membership, and vice versa. In a two group situation, membership function for group G_2 would be defined as 1 – membership value of group G_1.

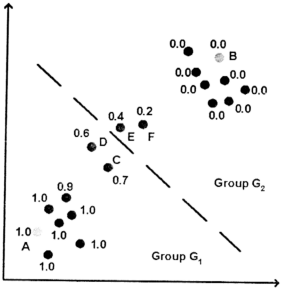


Figure 2.56 - Fuzzy Clustering Methods: The Number Alongside Each Point is its Membership Value with Respect to Group G_1.

In Figure 2.56, it can been seen that some points like A and B definitely belong to one or the other group. However, points such as C to F, appear to be intermediate between the two groups. A partition into two groups using the dashed line would disregard this information. It would simply state that A,C and D all belong to group G_1 without giving any idea of their relative strengths *of their membership*' of that group.

Common clustering methods are C-means or K-means clustering, fuzzy C-means clustering, mountain clustering method and subtractive clustering. K-means clustering or C-means clustering has been applied to a variety of areas, including image and speech compression, data preprocessing for system modeling and task decomposition in heterogeneous NN architectures [125]. Mountain clustering method was proposed by Yager *et. al.* in 1994 [125]. It is a simple and effective approach to approximate estimation of cluster centers on the basic of a density measure called the mountain function [125]. In subtractive clustering by Chiu *et. al.* [125], data points are considered as candidate for cluster centers [125].

Fuzzy C-means clustering (FCM) is also known as fuzzy ISODATA [125]. Fuzzy clustering techniques are used to analysis data when the membership functions have to be constructed from a given set of data [128,125]. It was proposed by Bezdek in 1973 as an improvement to the hard C-means (HCM) clustering technique [125]. The technique developed by Dezdek is described below [125,129,120,124].

FCM partitions a collection of n vector x_i, $i = 1, \ldots, n$ into a c fuzzy group. It finds a cluster center in each group such that a cost function of dissimilarity measure is minimized. FCM and HCM differs that FCM uses fuzzy partitioning so that a given point can belong to several groups with the degree of belongingness specified by the membership grades between 0 and 1 [125,127] .

The membership matrix U is allowed to have elements with values between 0 and 1 to accommodate fuzzy partitioning. Normalization is imposed to stipulate that the summation of degrees of belongingness for a data set is always equal to unity [125] [127].

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots, n \qquad \text{Equation 2.21}$$

The cost function for FCM is then a generalization of Equation 2.21 [125].

$$J(U, c_1, \ldots c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} \qquad \text{Equation 2.22}$$

where

u_{ij} is between 0 and 1

c_i is the cluster center of fuzzy group i

$d_{ij} = \| c_i - x_j \|$ is the Euclidean distance between ith cluster center and jth data point

$m \in [1, \infty)$ is a weighting component.

The necessary condition for Equation 2.22 to reach a minimum can be found by forming a new object function $\Im$ as following [125]:

$$\Im(U, c_1,...c_c, \lambda_1,...\lambda_n) = J(U, c_1,...c_c) + \sum_{j=1}^{n} \lambda_i \left(\sum_{i=1}^{c} u_{ij} - 1\right)$$
$$= \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^m d_{ij}^2 + \sum_{j=1}^{n} \lambda_i \left(\sum_{i=1}^{c} u_{ij} - 1\right)$$

Equation 2.23

where λ_j, $j = 1, \ldots, n$ are the Lagrange multiplier for the n constraints in Equation 2.21. By differentiation $\Im$ with respect with to all its input arguments, the necessary conditions for Equation 2.22 to reach its minimum are [125] :-

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

Equation 2.24

and

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$$

Equation 2.25

The FCM is an iterated procedure through the preceding two conditions. In batch mode, FCM determines the cluster center c_i and the membership matrix U using the following steps [110] :-

1. Initialize the membership matrix U with random values between 0 and 1 so that the constrains in Equation 2.21 are satisfied.
2. Calculate c fuzzy cluster center c_i, $i = 1, ... c$, using Equation 2.24
3. Compute the cost function according to Equation 2.22. Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.
4. Compute a new U using Equation 2.25. Go to step 2.

The cluster center can also be first initialized followed by the iterative procedure. FCM is not guaranteed to converge to an optimum solution. The performance depends on the initial cluster center, allowing the use of another fast algorithm to determine the initial cluster center or to run FCM several times, each starting with a different set of initial cluster centers [125]. FCM is used in medical segmentation and qualitative modeling [125].

2.8 CHAPTER SUMMARY

Chapter 2 has covered the background knowledge for the techniques and technologies used in this research. It begins by giving a brief introduction to digital image processing. The next section covers the MRI technology, exploring various issues about the images used for this research. The following section cover the 3D imaging pipeline that illustrates the steps taken to process medical image into obtain 3D models. Various biomedical image segmentation techniques were discussed next. In this section, segmentation validation issues are covered along with segmentation applications. After the human femur was illustrated, previous segmentation research done using MR bone image was explained. The next two topics are about the research techniques : artificial neural networks and fuzzy logic.